How neural is a neural net?

Bio-inspired computational models and their impact on the multiple realization debate

by

Mahi Hardalupas

MSci, University of Bristol, 2015

Submitted to the Graduate Faculty of the

Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2021

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation is presented

by

Mahi Hardalupas

It was defended on

April 6, 2021

and approved by

David Danks, Department of Philosophy, Carnegie Mellon University

Sandra Mitchell, Department of History and Philosophy of Science, University of Pittsburgh

David Plaut, Department of Psychology, Carnegie Mellon University

Dissertation co-director: Colin Allen, Department of History and Philosophy of Science, University of Pittsburgh

Dissertation co-director: Mazviita Chirimuuta, Department of Philosophy, University of Edinburgh

How neural is a neural net?

Bio-inspired computational models and their impact on the multiple realization debate

Mahi Hardalupas, PhD

University of Pittsburgh, 2021

My dissertation introduces a new account of multiple realization called 'engineered multiple realization' and applies it to cases of artificial intelligence research in computational neuroscience. Multiple realization has had an illustrious philosophical history, where multiple realization is when a higher-level (psychological) kind can be realized by several different lower-level (physical) kinds. There are two threads in the multiple realization literature: one situated in philosophy of mind and the other in philosophy of science. In philosophy of mind, multiple realization is typically seen as arbitrating a debate between metaphysical accounts of the mind, namely functionalism and identity theory. Philosophers of science look to how multiple realization is connected to scientific practice, but many question what it is useful for outside of philosophy of mind.

My dissertation addresses this gap by drawing on cases from machine learning and computational neuroscience to show there is a useful form of multiple realization based on engineering practice. It differs from previous accounts in three ways. First, it reintroduces the link between engineering and multiple realization, which has been mostly neglected in current debates. Second, it is explicitly perspectival, where what counts as multiple realization depends on your perspective. Third, it locates the utility of engineered multiple realization in its ability to support constraint-based reasoning in science. This provides an answer to concerns about the utility of multiple realization in the philosophy of science literature and explains how deep neural networks

can provide understanding of the brain. The first half of this dissertation proposes my account of Engineered Multiple Realization and applies it to scientific cases. The second half considers implications and connections to the modelling literature.

Table of contents

List of figures

Preface

I never would have predicted that I would be finishing this document and ending my time in Pittsburgh in the middle of a pandemic. In these circumstances, I am especially grateful and indebted to several people who made the writing and completion of this dissertation both possible and enjoyable.

First of all, I would like to thank my advisors, Colin Allen and Mazviita Chirimuuta. From the start, they encouraged me to pursue a project I was passionate about and helped chisel away at several rough drafts of this dissertation until it could reach the more sculpted final version you see before you. I feel incredibly lucky to have had the opportunity to work with them while they were both in Pittsburgh.

I also would like to sincerely thank my committee for their support and guidance throughout this project. Thank you to David Danks, for pushing me to carefully consider my audience and how to frame this dissertation; to Sandy Mitchell, for raising pertinent objections from a biological perspective and helping me sharpen my arguments; and to Dave Plaut, for enduring my (hopefully not too tedious) questioning on the purposes of biological plausibility and teaching me about the art and science of computational modelling.

The HPS graduate student community has been invaluable as a source of academic support and inspiration over the years, as well as much of what has made my time in Pittsburgh particularly memorable! Thanks to Nuhu Osman Attah, Trey Boone, David Colaço, Katie Creel, Haixin Dang, Marina DiMarco, JP Gamboa, Dana Matthiessen, Joe McCaffrey, Jacob Neal, Marina Baldissera Pacchetti, Evan Pence, Willy Penn, Lauren Ross, Zina Ward, Tom Wysocki, and several others. A special thank you goes to Siska De Baerdemaeker, Morgan Thompson, Nedah Nemati, and

Annika Froese, for their friendship and for helping me see a path through difficult times as well as being there to celebrate good times. I could not imagine my time in Pittsburgh without them!

Several people at Pittsburgh have been instrumental in keeping things running. A big thank you to the administrative staff during my time here: Natalie Schweninger, Kathy Rivets, Katie Labuda, Matt Ceraso, Diana Volkar, and Joann McIntyre, who sadly passed away just as I was finishing this dissertation. Thank you also to the staff of the Center for Philosophy of Science: Joyce McDonald (especially for all the help with the NeuroTech conference!), Carolyn Oblak, Alex Magee, and Cheryl Greer.

Another round of thanks is definitely needed for those beyond the world of Pitt HPS and Philosophy. Thank you to all my roommates over the past 6 years! Thank you to several friends on this side of the Atlantic and across the pond (and sometimes both!): Andrea Valentino, Dig, Gjeta Gjyshinca, Jasper Bone, Johanna Preston, Lydia Snodin, Martin Jonak, Nikianna Dinenis, Ray Schuur, Tolani Olonisakin, Rony Patel, Zhanna Budenkova, Sam & Julia Speight, Zoe Hunter-Gordon, Pierre Gianferrara, Marilena and Alexa Netty. Thank you to my godfather Alex Taylor and his family, for their support over the years. Thanks also to Carmela and Tony, as well as the rest of the White and De Mola families!

Thank you to my family: my parents, Yannis Hardalupas and Evi Simigdala-Hardalupas, and to my brother, Phil-Simon Hardalupas, for their unquestioning support throughout this PhD and my life, as well as to Evangelia Chardaloupa, Marika Simigdala, and Lena Hardalupas. I am incredibly lucky to have them in my corner. A final word of thanks to Davide De Mola, for having more faith in me than I have, for always knowing how to put a smile on my face, and for everything else, including the proofreading!

1.0 Introduction

The brain is like an oven, hot and dry,
Which bakes all sorts of fancies, low and high.
The thoughts are wood, which motion sets on fire;
The tongue a peel; the hand which draws, desire.
By thinking much, the brain too hot will grow
And burn them up; if cold, fancies are dough.
~ Margaret Cavendish, Nature's Oven

Biology and engineering have long drawn on each other for mutual inspiration. In early mythologies, artificial systems are often based on nature, such as the mechanical eagle created by Hephaestus to punish Prometheus for giving humans the gift of fire (Cave & Dihal, 2018; Mayor, 2018). Actual automata followed suit, with peacock automata built by al-Jazari in the 13th century, and a defecating duck automaton built by Vaucanson in the 18th century (Franchi & Güzeldere, 2005; Riskin, 2016). In the present day, we have robots built to mimic rodents, such as Shrewbot (Mitchinson et al., 2011) and bio-inspired materials that mimic the iridescent properties of butterfly wings (Saito, 2011). Not only do biological systems inspire artificial systems; they are frequently compared and contrasted to them. Neuroscience is no exception to this trend, in fact, it is replete with examples. Just as Cavendish compared the brain to an oven, Du-Bois Reymond paralleled the brain to the telegraph (Otis, 2001), Sherrington wove analogies between the brain and a Jacquard loom (Sherrington, 1940), and the comparison between brain and computer has persisted to the present day.

However, despite the enduring allure of understanding the brain in terms of a machine and vice-versa, the question of the utility of artificial systems in scientific research remains a fraught one. One reason for this is the tension between engineering and basic science research. An engineer

designs systems to use in the world, while a scientist develops new technologies to gain greater understanding of the world. A scientist defines her success in terms of attaining understanding of *how* and *why* a system works. An engineer defines her success in terms of whether the system works at all *in practice*. On this view, as products of engineering, artificial systems do not clearly inform the realm of basic scientific investigation. It is basic science that informs engineering, not the other way around. In this dissertation, I bridge this distinction by exploring the utility of Artificial Intelligence (AI) systems for scientific research in neuroscience.

1.1 Artificial Intelligence

The purported success of AI as a research program to illuminate our understanding of cognition has waxed and waned over the years. Different approaches have emerged as dominant at different times but two important ones for the purposes of my narrative are classical computationalism and connectionism. From cybernetics, McCulloch & Pitt's 1943 paper 'A Logical Calculus of the Ideas Immanent in Nervous Activity' is sometimes considered to be the first computational theory of mind and brain (Piccinini, 2004).[1] As Abraham (2019) notes, since its inception, AI research was being continuously assessed with respect to its biological similarity and, often, its connection to neurophysiology. This neurophysiological focus was discarded in the dominant paradigm of modelling cognition in the 1950s and 60s—classic computationalism, or what later came to be known as GOFAI: 'Good Old Fashioned Artificial Intelligence' by

[1] Though, arguably some would consider the history of computational approaches to thought (in some form) to start much earlier, such as in the work of Hobbes and Leibniz, or with respect to connectionist ideas, the work of Herbert Spencer and William James (Isaac, 2019; Medler, 1998).

Haugeland (1989). The classical computationalist approach rested on the assumption that cognition could be achieved through manipulating localist symbolic representations using well-defined rules. While neural network research had continued during these decades, it was not until the 1980s, when Parallel Distributed Processing (PDP) emerged as the iconic connectionist approach and became a dominant approach to modelling cognition, that neurophysiological concerns regained status in assessing AI. In contrast to classical computationalism, PDP models interpreted cognition as arising out of parallel processing of sub-symbolic elements resulting in distributed representations. When defending their approach, PDP modellers argued that their models corresponded more closely to how the brain actually works (McClelland et al., 1987)—a factor that was not deemed as important in classical computationalism.

Ultimately, concerns about the ability of AI to drive progress both in computer science and cognitive science meant it was subject to funding cuts, but recently there has been an AI renaissance. In the 2010s, the convergence of a trio of factors—the ease of procuring large, labelled datasets from internet users, the advent of powerful GPU-based (Graphics Processing Unit) computing systems, and technical advances—paved the way for a resurgence of interest in neural network approaches to AI in the form of Deep Learning. Deep Learning refers to methods of training "deep neural networks", multi-layer networks of nodes with more than one hidden layer, in order to succeed at certain tasks. The success of deep learning approaches led to a trend in considering how AI could be used in all sorts of contexts, from game-playing, machine translation, calculating the risk of recidivism, and even to scientific discovery.

In 2012, the success of AlexNet, a deep convolutional neural network, in the ImageNet object classification challenge solidified the promise of deep learning in domains of interest to neuroscience, such as computer vision. Much of the progress in machine learning has been

divorced from any commitment to try and model the brain, which was more important in the original PDP approach (Stinson, 2020). However, sparked by the success of AlexNet, there has been a wave of interest in exploring how a combined effort by AI researchers and neuroscientists might be of mutual benefit to both research programs. AI researchers, such as Geoffrey Hinton, Demis Hassabis, and Yoshua Bengio, argued there was much to learn from using the brain as a source of inspiration for producing better-performing AI (Bengio et al., 2015; Hassabis et al., 2017; Sabour, Frosst, & Hinton, 2017). Meanwhile, several neuroscientists became interested in how these networks could advance understanding of cognition and the brain (Guest & Love, 2019; Kriegeskorte, 2015; Lindsay & Miller, 2017; Richards et al., 2019; Yamins & DiCarlo, 2016). These researchers will often appeal to the historical connections between the fields of AI and neuroscience, as well as make claims about how artificial neural networks are designed and produced based on their biological counterparts. However, there is also deep disagreement over the extent of the purported similarity between artificial neural networks and the brain: disagreement exists on whether the learning algorithms used (e.g. back-propagation) are biologically plausible; whether supervised learning is analogous to human learning; and whether the architecture of artificial neural networks mirrors the hierarchical architecture of biological neural networks.

This research raises a panoply of questions: how can biologically plausible AI help neuroscientists understand the brain? What does it even mean for AI to be biologically plausible? In what sense are these engineered systems models of the brain or cognition? Are the capacities of AI systems really that similar to human capacities? My dissertation seeks to explore, interrogate and maybe even answer some of these questions.

1.2 Multiple realization

To frame my discussion of aforementioned issues and questions, I adopt the lens of the multiple realizability debate. Multiple realizability is the thesis that "higher-level" psychological kinds could be realized by several different "lower-level" physical kinds or, in other words, the same state or process could be realized in different ways. While the original debate relied on thought experiments involving Martians and Swiss cheese brains, nowadays philosophers are more interested in considering actual examples from the sciences. This can be construed as an 'empirical turn' that shifted the debate away from *multiple realizability*, which dealt with the conceptual possibility of kinds or processes being multiply realized, to *multiple realization*, which deals with whether kinds or processes are actually multiply realized. It is this latter debate surrounding multiple realization which I focus on in this dissertation.

However, in the midst of this empirical turn, most contemporary discussions of multiple realization do not substantially engage with examples of engineered systems—some examples from synthetic biology being a notable exception (Koskinen, 2019). This is surprising given the origins of the philosophical debate about multiple realization were influenced by developments in the field of AI at the time (Fodor & Block, 1972; Putnam, 1967). Two justifications for this oversight can be found in the literature. First, there is an impression that engineered cases are not advanced enough to count as actual cases of multiple realization of cognitive processes rather than as thought experiments in multiple realizability arguments (Polger & Shapiro, 2016, p.147). Since the 'empirical turn' has led philosophers away from assessing thought experiments for multiple realizability, then this would explain why engineered cases are dismissed. Second, there is a view that examples of multiple realization among artifacts and engineered systems are trivial and thus do not inform scientific inquiry (Polger & Shapiro, 2016; Preston, 2009). Thus, if we are interested

in multiple realization in order to inform philosophy of mind and philosophy of science debates, then it is not clear what we can expect to gain from discussing purportedly trivial artificial cases. I think dismissal of engineered cases on these grounds is misguided. On the one hand, the historical and current interest in AI, as applied to cognitive science and neuroscience, suggest that at least some scientists consider engineered systems to be advanced enough to inform scientific enterprises and generate understanding. On the other hand, to consider all engineered cases as trivial ignores the challenges that face scientists and engineers who attempt to reproduce various functions, such as the engineering of planes that could fly. For these reasons, I argue it is high time we reconsidered how multiple realization applies to engineered cases.

In doing so, I can also re-evaluate other common assumptions made about multiple realization. The original motivation of the multiple realization debate was to arbitrate debates on functionalism in philosophy of mind, where multiple realization supported functionalist accounts of the mind and the lack of multiple realization indicated support for reductionist accounts of the mind. This common framing of multiple realization as connected to reductionism and anti-reductionism obscures other roles that multiple realization can play in philosophical debates. By revisiting the utility of multiple realization in engineering contexts, I identify a role it can play in constraint-based reasoning, where engineering artificial systems informs scientists about the constraints that affect functions and realizers of a function.

1.3 Dissertation preview

This dissertation is composed of four substantive chapters. The first develops the core conceptual framework of my account of engineered multiple realization (EMR). The rest of the

chapters illuminate different facets of this account in applying it to examples from computational neuroscience.

In the first chapter, I introduce and explicate my account of EMR. In particular, I emphasise three central claims of my account. First, multiple realization connects to an important part of scientific practice: the engineering of artificial systems. Second, an explicitly perspectival account of multiple realization is needed to understand the engineering of artificial systems for scientific purposes. On this view, the assessment of multiple realization must be done from a particular perspective that depends on the goals, methods, and definitions of biological similarity assumed by researchers. Third, EMR provides a philosophical framework for understanding constraint-based reasoning in science.

In the second chapter, I apply my account of EMR to two cases drawn from contemporary neural network research in neuroscience. These cases are examples of two different approaches to using artificial neural networks in computational neuroscience, which differ on what counts as multiple realization: the anatomy-driven and the performance-driven approach. I show that these correspond to different perspectives and hence different forms of multiple realization. Despite these differences, I explain how both approaches contribute to constraint-based reasoning, which aims to explain how common features of a group of systems arise with reference to the shared set of constraints these systems are subject to. This demonstrates the utility of constraint-based reasoning as a new role multiple realization can play in philosophy of science.

In the third chapter, I argue for interpreting Deep Neural Networks (DNNs) as artificial 'Krogh organisms' (Green et al., 2018). Specifically, this chapter seeks to answer in what sense DNNs are a model of the brain. One proposal made by neuroscientists is that DNNs are artificial model organisms. I assess this view by evaluating DNNs with respect to extant philosophical

accounts of animal model organisms and argue that, while DNNs bear some similarity to model organisms, current DNNs are more accurately interpreted as artificial Krogh organisms. I connect these uses to the two methodological strategies used in computational neuroscience research that I identify in the second chapter. Then, I connect this to my framework for engineered multiple realization by explaining how and when these strategies contribute to constraint-based inferences. Furthermore, I end by defending my view against other accounts of DNNs as models of the brain by commenting on the similarities and differences between machine learning models and animal models.

In the fourth chapter, I defend my account against mechanistic interpretations of DNNs in computational neuroscience. I show how assumptions about the unlikelihood of multiple realization underlie arguments for interpreting DNNs as mechanistic models. Using my account of EMR, I argue that these assumptions currently fail to justify the mechanists' arguments. In particular, I present an underdetermination argument that can be made for current DNN research that undermines the possibility of drawing conclusions about mechanistic similarity between the brain and DNNs based on current neural network research.

To conclude, I summarise some key takeaways from this project and suggest some future directions.

2.0 Engineered multiple realization, constraint-based inference and computational neuroscience

> *Although I am not pressing for a neurophysiological approach to AI, I am unlike many AI people in that I believe that any AI model eventually has to converge to brainlike hardware, or at least to an architecture that at some level of abstraction is "isomorphic" to brain architecture (also at some level of abstraction) […] I believe that the level at which the isomorphism must apply will turn out to considerably lower than (I think) most AI people believe*
> ~ Hofstadter (1985), p.644

In this chapter, I provide a new account of multiple realization - engineered multiple realization - that centres a largely neglected source of multiple realization: engineered artificial systems. I argue my account has three advantages over extant accounts of multiple realization: it recognises the role of engineering in scientific practice, it is an explicitly perspectival account, and provides an answer to the question: what can multiple realization do for philosophy of science? Regarding this last question, I propose that engineered multiple realization provides a philosophical framework for constraint-based reasoning in science.

2.1 Introduction

Imagine you are on a journey in a forest and you happen upon an extra-terrestrial creature that is totally unlike anything you've ever seen or read about before. The extra-terrestrial is motionless, and you wonder if it's hurt and try to figure out if it's in pain. You recall a podcast you listened to recently about the latest scientific research on pain that discussed the neural structures underlying how humans feel pain. As it happens, the extra-terrestrial creature has a translucent

body so you can see the organs inside. However, from what you can see, this creature doesn't have anything that looks like these neural structures. So, could it still feel pain?

This question is at the heart of the debate on multiple realizability. Multiple realizability is the thesis that "higher-level" psychological kinds could be realized by several different "lower-level" physical kinds[2] or, in other words, the same state or process could be realized in different ways. If multiple realizability is true, it would be possible that both an extra-terrestrial creature and I feel pain even if we have very different physical structures (Putnam, 1967). The multiple realizability thesis originated in philosophy of mind where its importance lay in its ability to arbitrate the debate between functionalism and reductionism about mental states. Functionalism is a theory postulating that a mental process is determined by its inputs, outputs, the relations between them, and the functional role it plays in a system. This is opposed to reductionism (or identity theory) where mental processes reduce to (or are identical to) neural processes. Multiple realizability is thought to support functionalism since if a mental process could be realized in lots of different ways, then it can't depend on specific neural details. In other words, if pain is multiply realized, I wouldn't be able to deny the extra-terrestrial feels pain just on the basis of its lack of specific physical structures.

However, there are plenty of real-life examples closer to home than my imagined extra-terrestrial. Does an octopus feel pain? How about a fruit fly? Over the past decades, philosophers have shifted to evaluating the multiple realizability debate with real-life scientific cases instead of hypothetical ones (Aizawa & Gillett, 2009; Bechtel & Mundale, 1999; Figdor, 2010; Polger & Shapiro, 2016). Thus, the debate has shifted away from multiple *realizability*, the conceptual

[2] Here levels talk is in quotation marks to reflect the growing discontent with levels terminology as discussed in Potochnik & Mcgill (2012). However, multiple realizability can also be based on an egalitarian ontology, as in Piccinini (2020).

possibility that mental processes *could* be realized in different ways, to multiple *realization*, whether mental processes *are*, in fact, realized in different ways. Despite this shift, multiple realization is still frequently interpreted solely through the lens of the debate between functionalism and reductionism where evidence for the multiple realization thesis is taken to be good news for functionalism and bad news for reductionism.[3] I argue this prevailing framing of the issue fails to consider other roles multiple realization could have beyond arbitrating the functionalism-reductionism debate.

Looking beyond the functionalism-reductionism debate, a natural place to look for what multiple realization (henceforth MR) can do is to look to how MR features in scientific practice. However, many philosophers claim that MR loses its force when transferred from philosophy of mind to philosophy of science. The concept of multiple realization formulated in philosophy of mind was based on a syntactic view of scientific theories, which defined a scientific theory in terms of theorems and emphasised the need to translate theories into axioms. Given the widespread rejection of the syntactic view in favour of alternatives such as the semantic view in philosophy of science, Klein (2013) argues that debates around MR can no longer be motivated. Kaplan (2017) questions its relevance to mechanistic explanation and Michel (2018) argues the multiple realization argument cannot settle debates on the existence of pain in fish. Both of these arguments cast doubt on the utility of multiple realization to debates in philosophy of cognitive science more generally. Miłkowski (2018) memorably summarises the struggle to find a role for multiple realization, "I still have to see why this notion [MR] is respectable in current scientific metaphysics. It does not help in the debate between embodied cognition and computationalism. It

[3]For example, Polger and Shapiro (2016), in questioning the prevalence of multiple realization, endorse a modest identity theory that they argue preserves the autonomy of psychology (see chapter 10).

is useless in arguments against reductionism. So what is it good for?" (p.361). There is indicative of the strong impression within philosophy of science that multiple realization is a solution in search of a problem and, at best, should be confined to the tomes of metaphysics where it originated.

In this paper, I argue for an account of multiple realization that locates its utility in providing a philosophical framework for constraint-based reasoning in science. While I am by no means the first to defend the role multiple realization can play in philosophy of science (Batterman, 2000; Aizawa, 2018; Chirimuuta, 2018), I contend that to rehabilitate multiple realization in philosophy of science, we need to revisit the link between engineering and multiple realization. This was a salient theme in early multiple realization papers (Fodor & Block, 1972; Putnam, 1967, 1975), full of computers that could feel pain or living robots, but these artificial examples have been mostly neglected in the current debate.[4] My account of Engineered Multiple Realization (EMR) differs from other recent accounts in three ways: by returning to artificial examples drawing on cases in computational neuroscience where MR is "engineered", by being an explicitly perspectival account of MR, and by locating MR's utility in constraint-based reasoning in science.

In section 2.2, I provide a definition of engineered multiple realization (EMR) to describe cases where scientists attempt to create multiple realizations by engineering artificial systems and motivate why we should incorporate engineered cases into our accounts of multiple realization. In section 2.3, I will outline how EMR is an explicitly perspectival account of multiple realization and spell out the conditions for a perspective from which EMR is assessed. In section 2.4, I explain

[4]A notable exception is Koskinen (2019) who uses cases from synthetic biology to argue that multiple realizability is a desirable design heuristic for biological engineering. Figdor (2010) also points out this oversight: "since neuroscience in general has been the primary source of evidence used in "empirically based" arguments against MR, this [paper] obeys a de facto restriction to an "evolved-biological" debate. A truly empirically based debate would also include cognitive systems we can engineer, biologically or artificially" (p.420).

what this account of EMR is useful for by discussing how it supports constraint-based reasoning through two strategies. Finally, in section 2.5, I reflect on the benefits of EMR in contrast to traditional approaches to multiple realization. I conclude that my new EMR framework demonstrates the utility of multiple realization to philosophy of science and can pave a route to understanding how computational models such as Deep Neural Networks (DNNs) inform cognitive science.

2.2 Motivating Engineered Multiple Realization

EMR occurs when an agent[5] engineers a system S that performs a function F in a way that is different from F's instantiation in a target system T. For example, birds, planes and helicopters all instantiate the function, F, of flight but their methods of flight differ substantially. In EMR, whether the artificial system instantiates the same function in relevantly different ways is open to investigation. Is the helicopter instantiating flight in the same way as a bird? As we will see later, different norms can be used to evaluate whether an engineered system counts as *really* instantiating the same function in a different way. While it is often acknowledged that MR is prevalent in artificial systems, there is either little discussion of how this connects with scientific enterprises or it is treated as a trivial form of MR (Polger & Shapiro, 2016; Preston, 2009). This is at odds with one of the motivating examples in papers originating the concept of MR: whether artificial systems, specifically Turing machines, could realize psychological kinds (Fodor & Block, 1972; Putnam, 1967). A re-evaluation of the connection between MR and artificial systems is long

[5] 'Agent' here is ambiguous between individual agents and groups of agents, with the latter being more common in scientific research.

overdue. Not least because, as I will emphasise in this section, EMR has long been an important part of scientific research in order to provide understanding of the target system, T, or the function, F.

Consider the eighteenth-century fascination with the construction of automata, mechanical devices that imitated human beings or animals. While often dismissed as more artistic than scientific enterprises, some automata played an important role in scientific reasoning prompting questions about how these automata worked and whether their operation was similar to humans. A famous example is Jacques Vaucanson's flute-playing automaton that "breathed", delighting audiences in Paris, and impressing Voltaire, who declared Vaucanson to be "Prometheus's rival" (Riskin, 2016, p.120). At the time, this flute-playing automaton piqued the interest of Condorcet who explained how the Academy of Science was tasked with examining the automaton and concluded that its mechanism relied on the same operations as a human flautist (Canguilhem, 1961). Here is a case of EMR where scientists interpreted the automaton as performing the same function F of flute-playing as the human "target system" T. This example also serves to demonstrate the scientific goals bound up in EMR. Riskin (2007) argues that the development of automata in the eighteenth century was linked to a pervasive desire for physiological correctness demonstrated by the interest in the similarity between the operation of artificial systems and natural (human) systems. Vaucanson harboured grand scientific ambitions for understanding human processes when he outlined his plan "to create an automatic figure whose motions will be an imitation of all animal operations [in order to] be able to carry out experiments on animal functions, and […] draw conclusions from them which will allow us to recognize the different states of human

health" (as quoted in Riskin (2007), p.260).[6] The process of generating these EMRs was motivated by scientific goals to understand human functions and processes better.

The same motivation is found in early cybernetics research where EMRs were artificial systems meant to engender understanding of the brain such as British cybernetician William Grey Walter's robot tortoises, simple robots with a two-cell "nervous system" that could scaffold our understanding of human nervous systems (Holland, 2003).[7] In a similar way, contemporary researchers use deep neural networks (DNNs) to serve as models of brain function. DNNs are networks of nodes organised in multiple layers that "learn" relationships between inputs and outputs using algorithms such as backpropagation. In 2012, the success of AlexNet, a DNN, in visual recognition tasks sparked enthusiasm amongst neuroscientists and artificial intelligence researchers alike about the potential of deep learning to "revolutionise" our understanding of the mind and brain (Krizhevsky et al., 2012). Typically seen as crucial for this success in the domain of vision is the use of a particular kind of DNN called a convolutional neural network. Convolutional neural networks (CNNs) have additional architectural features that are better suited to extracting features that are invariant across transformations. For example, CNNs are more able to identify images of the same cat as a cat even if the cat is rotated or in a different part of the image.

Both philosophers and scientists have claimed that the success of these models lies in the networks' similarities to biological systems themselves (Buckner, 2018; Kriegeskorte, 2015),

[6] This point is corroborated by Canguilhem (1963) where he references a similar conclusion drawn by Doyon and Liaigre about the link between medical research and the construction of mechanical automata (p.510).

[7] For later examples, see also the motivation for Fukushima's (1980) Neocognitron work and discussion of this in neurorobotics (Mitchinson et al., 2011).

prompting debate over the extent to which these systems actually work in a similar way to humans and other animals and whether they can inform our understanding of the brain. I contend that this is an important goal of multiple realization - by engineering artificial versions of biological systems, scientists aim to use them to better understand the biological systems themselves.[8] This commits researchers generating EMRs to two assumptions:

a) Engineered system S is able to reproduce the function F[9] of a target system T,

b) The process of engineering S will provide greater understanding of the target system T or the reproduced function F.

These two assumptions will need more unpacking to fully appreciate the commitments of EMR. However, you may immediately challenge a) and question whether it is possible to claim that the engineered system S is reproducing the function F. What does it mean to say a DNN is "really" doing object recognition? Such skepticism is well-founded and in the next chapter, I will go into more detail regarding how these attributions of functions have been challenged and the strategies scientists use to deal with this. For now, it suffices to say that at least some scientists do take these artificial systems to be doing the same thing as a target system rather than treating them as simulations. This is an important distinction because simulations do not have the same functional capacities of their target system – to paraphrase Dennett (2014), a simulation of a hurricane is not itself a hurricane. In contrast, a case of EMR must be interpreted as sharing at least

[8] While I focus mainly on EMR as it occurs in computational cognitive science, it is also common in other sciences such as synthetic biology (Koskinen, 2019) and in applied physics, for example, when researchers create artificial materials that exhibit optical properties found in biological systems (Kawamura et al., 2016).

[9] By 'function' here, I mean some form of biological function, rather than a mathematical function or input-output relation that could be reproduced in both human brains and computational systems. I leave it open how one wants to interpret biological function here though whichever account one chooses would have to apply to both biological and artificial systems.

some of the same functional capacities as the target system. Vaucanson's flute-playing automaton interested scientists not because it was a simulation of flute-playing but because scientists took it to be an instantiation of flute-playing. Similarly, computational neuroscience assumes that computers and the brain are doing the same thing: information processing. This is clear in early cybernetics[10] and artificial intelligence research by Simon and Newell,[11] who explicitly claim "the thinking human being is also an information processor" (Simon & Newell, 1964, p.281). Later, in the 1980s, Parallel Distributed Processing models, a precursor to today's deep learning models, were also interpreted as representative of human's natural information-processing capacities (McClelland et al., 1987). In the present day, computational neuroscientists consider DNNs good models for the brain because they are able to solve the same behavioural tasks as humans and animals such as being able to correctly identify objects in images (Yamins & DiCarlo, 2016). Since these cases are interpreted as performing the same function, they fulfil the criteria for what constitutes an EMR.

As I have defined it, EMR occurs when an agent engineers a system S that performs a function F in a way that is different from F's instantiation in a target system T. In this section, I motivated the first key feature of EMR by showing the importance of including engineered examples into an account of multiple realization. In the next two sections, I spell out the details of the remaining two distinctive features of EMR: first, its explicit perspectivism, and second, its utility in constraint-based reasoning in science.

[10] See Abraham (2019).

[11] See Dick (2015) for the history of this assumption in Newell and Simon's work.

2.3 EMR and perspectivism, or what counts as EMR

One key feature of my account is that it is an explicitly perspectival account, where what counts as making instantiations "different" enough for multiple realization depends on your perspective. Perspectivism[12] in philosophy of science recognises the situated nature of scientific knowledge such that we cannot separate the generation of scientific knowledge from its background context broadly understood (Danks, 2020; Giere, 2010; Massimi, 2017; Potochnik, 2012; Van Fraassen, 2008). Since scientific knowledge is situated in historical and cultural research practices, there will be no one overriding perspective that encompasses all relevant information about a phenomenon – that is, scientific representations will always be partial. This has implications for understanding scientific practice. Given the partial perspectival nature of scientific representation, it follows that the best description of science is scientific pluralism – the need for multiple theories, models or representations in order to understand and explain the world through science (Mitchell, 2020). This is especially true when building artificial systems where full emulation of the natural system is challenging, if not impossible. An engineer is forced to make choices about what aspects of a system must be simplified or abstracted, which will depend on what is of interest scientifically. This will often also be bound up with social and political implications of how our knowledge is represented (Adam, 1998). To claim that MR is perspectival is to recognise the contingency of our claims pertaining to MR and accept there are different norms governing what counts as successful MR.

[12] Sometimes in the philosophy of science literature, perspectivism is associated with 'perspectival realism' but I do not consider the realist implications of perspectivism in this paper.

Classic accounts of MR, such as Putnam's, treated MR as concerned with implantation in lower-level kinds. The standard way multiple realization would occur is if the same function was implemented in different ways in different substrates. Consider Putnam's remark that "[w]e could be made of Swiss cheese and it wouldn't matter" (Putnam, 1975, p.291). Or consider Pylyshyn's (1980) thought experiment of neurons being progressively replaced with silicon transistors that are identical in terms of input-output function. The presumption in these examples was that you would still go on functioning as before since the function was multiply realizable, meaning it could be successfully realised in different substrates. In more recent literature, there is a shift away from the view that MR is purely to do with substrate. For example, Millikan hints at the interest-dependence of MR:

> "Sometimes different mechanisms that accomplish the same [thing] operate in accordance with different principles; other times they represent merely different embodiments of the same principles. Or we might say, sometimes looking more closely at the mechanism helps to explain how it works; sometimes it reveals only what stuff it is made of. It is only the former kind of difference that makes interesting 'multiple realizability'"(Millikan, 1999, pp.61-2).

This is also reflected in Polger & Shapiro's (2016) account, which rejects the view that difference in substrate is enough to permit multiple realization; they suggest that corkscrews made of different materials will not be enough to constitute multiple realization, there must be differences in the corkscrews' mechanisms. The perspectival nature of MR is briefly discussed or implicit in other discussions but not forefronted as a central feature of MR (Miłkowski, 2018; Piccinini & Maley, 2014; Polger & Shapiro, 2016; Wimsatt, 1994). However, even if MR is discussed as interest-dependent, philosophers will often still discuss MR as if, once we've looked at scientific practice,

there is one form of multiple realization which eventually can be agreed upon. Thus, they assume that there is a fact of the matter about whether a kind or process is multiply realised or not.

Rather than accept this assumption, I argue we should embrace the perspectivism of multiple realization. In other words, multiple realization itself is multiply realized, as there are several different forms of MR depending on your perspective. Since MR was originally conceived as a counter to reductionism and identity theory, it was perhaps natural that it was focused on implementational details about substrate, which clearly mirrored candidate reductionist claims. However, without framing MR as in opposition to reductionism, we should move away from MR only being based on these kinds of differences. Once multiple realization is divorced from concerns about reductionism, this frees up space to reconsider its relation to scientific practice. In particular, I wish to emphasise that what counts as "biologically relevant" differences for MR is not static, as there is no one form of biological similarity that can override others when we try to understand cognition (Chirimuuta, 2020a). What counts as a relevant difference for MR is a matter of perspective. This aligns with how scientists are considering these issues as well. In their criticism of the current enthusiasm surrounding neural network research, Lake et al. (2017) criticise current notions of biological plausibility that are used to support neural network models as the best models of the brain. The crux of their criticism lies in how a narrow notion of biological similarity is used to assess DNNs' plausibility as models. For example, they often rely on stylised understandings of cellular neuroscience rather than systems neuroscience. Lake et al. point to the need for considering cognitive plausibility too, suggesting that we ought to treat models that are cognitively implausible with skepticism. In the same way, I argue that many different forms of biological similarity may constitute relevant differences for assessing EMR. Here is a suggestive but non-exhaustive list:

1) Cognitive/psychological similarity – are computational models and brains cognitively similar? (For example, do they exhibit the same patterns of behaviour?)
2) Neural similarity – are computational models and brains neurally similar?
3) Molecular similarity – are computational models and brains similar on the molecular level?
4) Algorithmic similarity – do computational models and brains use the same learning algorithm? (See, for example, debates about whether back-propagation is biologically plausible).
5) Representational similarity – do computational models and brains use the same representations? (See, for example, research comparing the representations used in machine learning to those in the brain).
6) Developmental similarity – is the process by which brains and computational models learn and adapt similar? (See, for example, debates about whether supervised learning is a biologically realistic process for human cognition).
7) Mechanistic similarity – do computational models and brains have similar mechanisms realizing cognitive processes?
8) Architectural similarity – do computational models and brains have similar architectures that support a similar representational hierarchy?
9) Ecological similarity – do computational models and brains engage in similar sorts of tasks? Do computational systems and brains need to be embodied in similar ways for comparison? (See, for example, debates in neurorobotics about whether a particular morphology is needed).

This should serve to demonstrate the tension in making claims about MR without considering what kinds of similarities or differences we take to be relevant for MR. Selecting one or more varieties of biological similarity from the above list will introduce different conceptions of MR.[13] Hence, why a perspectival account of MR is needed to make sense of scientific practice when engineering systems.

On my account, a perspective from which EMR is assessed depends on these three components:

a) A notion of biological similarity or relevance B,

b) The methodology or technique for modelling the system M,

c) The assumed goals of the scientists G.

Importantly, these components are variable and not set in stone, which counters the tendency in the literature to view multiple realization as something fixed that can be discovered about a particular kind. One could disagree that these components affect multiple realization so let me spell them out in more detail.

First, the choice of B, what is biologically similar or relevant to realizing a process affects whether something counts as multiply realized. If you decide that for an animal to exhibit pain, you care about psychological similarity, then you are setting a particular bar for what counts as multiple realization such as behaviourally similar features to humans experiencing pain. For example, Sneddon et al (2014) look for pain-avoidant behaviour and self-administration of

[13]For those who are concerned that this account of EMR is one that prioritises scientists' opinions on similar or different realizers, I contend my account need not place the authority of decisions on MR all to scientists. Even in cases where philosophers are the ones making decisions related to multiple realization, I still think we ought to endorse an account that allows for differences about what counts as important biological similarities according to philosophers and their own goals (metaphysical, ethical or otherwise) for MR. For example, the philosopher who wishes to use MR to argue against reductionism would need a concept of biological similarity based on implementation rather than on cognitive similarity. One limiting factor is that philosophers will often not have control over the methods and techniques used by scientists in engineering a candidate system S.

analgesics as a way to ‘test’ for animal pain. If you decide that what matters for identifying pain is neural similarity, then this commits you to a different form of assessing multiple realization such as by examining neural structures. For example, in discussing the possibility of fish pain, Key (2015) investigates whether fish have the appropriate neural machinery to feel pain. Here we see that decisions about what constitute biologically relevant differences can affect the threshold for what counts as realizing the same function or the features of the realizers scientists are interested in. The same occurs in computational cases. If a scientist defines navigation as being able to physically orient from a particular location to another, this commits them to certain forms of biological similarity that will influence what entities could realize navigation. For example, here we couldn’t use a non-embodied computer to realize this form of navigation, but an embodied robot could potentially realize it. What constitutes biological similarity and relevant differences will depend on the perspective.

The second component of an EMR perspective, M, concerns how the modelling approach affects the concept of biological similarity. Choices about B can affect appropriate methodologies, as was the case in the robot navigation example. However, choices about how to model a system also may determine the forms of biological similarity that can be investigated with the help of the model. This arises because the ability to translate or formalise biologically relevant features into a model will depend on what techniques and technologies are available. For example, using a predictive processing framework to model the brain could prioritise performance over anatomical features. Furthermore, the choice of B may commit a scientist to a set of biologically relevant features, but technological limits could prevent the scientist from being able to adequately incorporate these into their model. This reminds us once more that EMR must be temporally

qualified as it is dependent on what technology and methods happen to be available to scientists at a particular time.

Third, G recognises how scientists' goals shape what counts as successful EMR. A scientist's goals may override other information they know is relevant to modelling a phenomenon. For example, a computational neuroscientist using cognitive models can recognise that there are neural details relevant to understanding or explaining a cognitive process. However, their goal could be to have a tractable model of a phenomenon or to explain a process in psychological terms – this can motivate them to not include the neural detail and abstract away from it.[14] This decision, in turn, affects B, the set of features that constitute their working definition of biological similarity.

In summary, questions relating to EMR must be analysed through a perspectival framework in order to best capture scientific practice. What counts as successful multiple realization is ultimately perspective-dependent and temporally qualified. Since the components making up each perspective vary, we can't necessarily expect a consensus between different agents on whether a cognitive process is multiply realized or not. This question can only be asked from within a perspective.

2.4 EMR and constraint-based reasoning, or what EMR can do

The final key feature of my account is that EMR provides a philosophical framework for constraint-based reasoning in science. This addresses concerns about what MR can do for philosophy of science.

[14]See Chirimuuta (2020) for a recent argument discussing this in computational accounts of the brain.

The connection between constraints and MR has been made before (Shapiro, 2004; Towl, 2012), however it normally features in arguments against MR based on convergent evolution. Shapiro (2004) gives extensive treatment of this issue proposing that there will be few ways to realize human-like psychological capacities because human minds evolve under several constraints that limit the amount of phenotypic variation across realizers – what he terms the mutual constraint thesis (MCT). In other words, there will be restrictions on structure-function mappings such that we should expect only a few structures to be equipped to realize a function rather than many structures. There are reasons to challenge this view of biology, for example, Amundson (2000) provides several examples to argue against functional determinism – the view that statistically we can expect convergence to uniform design such that there will be few individuals with novel functional design. However, here I am particularly interested in a suggestion made by Weiskopf (2011) that Shapiro's discussion of constraints is perfectly compatible with multiple realization, since Shapiro's examples are of functional constraints, which will not determine the underlying physical mechanism. In agreement with Weiskopf (2011), rather than construe constraints as a threat to multiple realization, I argue they provide us with an answer for why scientists engage in EMR: it allows them to form and test hypotheses, and provide explanations in terms of constraints on systems.

Constraint-based reasoning aims to understand how the common features of a group of systems come about and does so by showing how these systems are all subject to a shared set of constraints (Green & Jones, 2016). This is a common explanatory strategy in biology – in order to understand why desert mammals might share common characteristics such as thin fur that is light in colour, scientists appeal to the shared set of environmental constraints such as living in high temperatures with little available water. These constraints make it desirable to have adaptations

that minimise water loss and aid in cooling the mammal down. This highlights important relations between the organizational or structural properties of a mammal and the functional properties they possess that are mediated by environmental constraints.

To expand on this, I will start by considering how to define constraints.[15] Ross (ms) provides four features that distinguish constraint-based explanations from standard explanations. In constraint-based explanations, "constraints are factors that

1) Limit the values of the explanatory target of interest,

2) Are often conceived of as separate from or external to the process they limit,

3) Are considered relatively fixed compared to other explanatory factors,

4) Structure or guide the explanandum outcome, rather than triggering it."

On this basis, she proposes a taxonomy of constraints: law-based constraints, mathematical constraints, and causal constraints. These are defined with reference to two conditions contrasting them to typical causal explanatory factors, which a) have a manipulable explanans and b) are based on empirical dependency relations. Causal constraints are those that satisfy both of these conditions, while law-based and mathematical constraints fail a) and b) respectively.

Both Shapiro (2004) and Ross (ms) have emphasised the limiting nature of constraints. However, there is another role of constraints which highlights their ability to afford or enable certain outcomes. Green & Jones (2016) recognise this dual nature of constraints, defining constraints as conditions that "both limit and afford a certain scope of possible structures and functions that can be instantiated in a system of a particular type" (p.345). For example, the sensitivity of skin limits our ability to withstand extreme temperatures but also affords tactile

[15] Note that this notion of constraint is different to that which is normally discussed in the philosophy of neuroscience literature, where constraints refer to constraining features on mechanistic explanations that reflect desiderata of a good explanation(Craver, 2007; Franklin-Hall, 2016; Illari, 2013).

discrimination. Similarly, the rigidity of an exoskeleton affords protection from the physical environment but limits the growth and flexibility of an organism. This positive role for constraints is also emphasised by the notion of 'enabling constraints' – those that promote particular outcomes in a system – discussed by Anderson (2015) and Raja & Anderson (2021). For example, in Raja and Anderson (2021), they argue that behaviour is an enabling constraint, where behaviour is not only an outcome of neural activity, but rather a constraint that also enables neural activity in the first place, by connecting it to the environment.

Based on these accounts, I build on Ross's taxonomy and present five broad, non-exhaustive, and potentially overlapping, categories of constraints in the literature:

(L) Law-like – constraints that are non-manipulable and based on scientific laws or principles (includes universal constraints (Shapiro, 2004), developmental constraints, and formal constraints (Green & Jones, 2016));

(M) Mathematical – constraints that are manipulable but not based on empirical dependencies (such as those that feature in topological explanations);

(C) Causal – constraints that are both manipulable and based on empirical dependencies thus fulfilling the conditions for causal explanation;

(H) Historical – constraints that are manipulable, based on empirical dependencies but are considered accidental or local constraints (these would be under L on Ross's taxonomy);

(E) Enabling – constraints that afford or promote certain outcomes (includes functional constraints (Weiskopf, 2011)).

These constraints are clearly relevant to engineering too. If I am designing an aquatic robot, the choice of a particular material may limit its flexibility but afford its robustness to changes of current in the water. In the context of modelling, we may impose formal constraints on our models

to capture or approximate a causal (C) or law-like (L) constraint that a biological system is also subject to. For example, the use of mathematical equations, like a stress-strain curve and corresponding Young's modulus, to capture the elastic material properties of the skin that can then be applied in a model. These formalised constraints can then be connected to possible design principles that both biological and non-biological systems instantiate.[16]

To gain insight into why scientists would generate EMRs, let's consider two things they could learn from EMR. The process of engineering a system S that performs F could:

1) Provide greater understanding of a target system T (that also performs F),

2) Provide greater understanding of reproduced function F.

I argue that each of these results rely on a corresponding strategy that utilises knowledge about constraints, as is represented in Figure 1. First, in order to gain greater understanding of a target system T, an iterative strategy is used which aims to make the engineered system S more similar to the target system T. Here the regulative ideal is to eliminate MR in the dimension(s) of interest. Through the iterative strategy, scientists will try to mirror constraints on the target system as a way to reduce MR in their engineered system. This relies on the assumption that increasing the number of constraints on their system is a way to limit the realizers of a function. Second, in order to gain greater understanding of the reproduced function F, a comparative strategy is used which looks to generate several different systems instantiating F. Here the regulative ideal is to maximise MR in the dimension(s) of interest. By generating greater variation between realizers, it allows scientists to better investigate possible enabling constraints that lead to particular outcomes, or test hypotheses regarding potential law-like constraints on a system.

[16] See Sterling & Laughlin (2015) for an example of this in the case of neuroscience.

Take object recognition as a candidate function. Defined broadly, it is the categorisation of objects into their correct groups. This definition introduces few constraints on an engineered system so that anything that successfully categorises objects into correct groups will realize the same function. Defined more narrowly, object recognition can be viewed as the categorisation of objects into groups as humans do it. This definition introduces more constraints such as behavioural constraints according to which engineered systems are expected to match human performance and the mistakes humans make. When the definition of biological similarity implies fewer constraints, the hope is that there will be radical MR where systems accomplish tasks of realized functions in very different ways. This would allow for the use of a comparative strategy, which can then inform understanding of the constraints involved in the function F. When the definition of biological similarity implies more constraints, the hope is that it is less likely there will be radical MR (though it is still possible depending on how crucial certain constraints are to affecting realization). This would allow for the use of an iterative strategy, which can then inform scientists' understanding of the target system T. In both cases, EMR explains why generating multiple realizations is useful: it allows us to reason about constraints on the target system, the function, or both.

In chapter 3, I'll spell out how these strategies and forms of EMRs actually work by introducing two approaches in computational neuroscience: a performance-driven approach, and an anatomy-driven approach. By applying EMR to these cases, I also highlight some methods used by scientists in pursuit of constraint-based reasoning. In chapter 4, I connect these strategies to certain accounts of models that are consistent with interpreting DNNs in terms of artificial Krogh organisms or model organisms. For now, I end by reflecting on some benefits of my account of EMR.

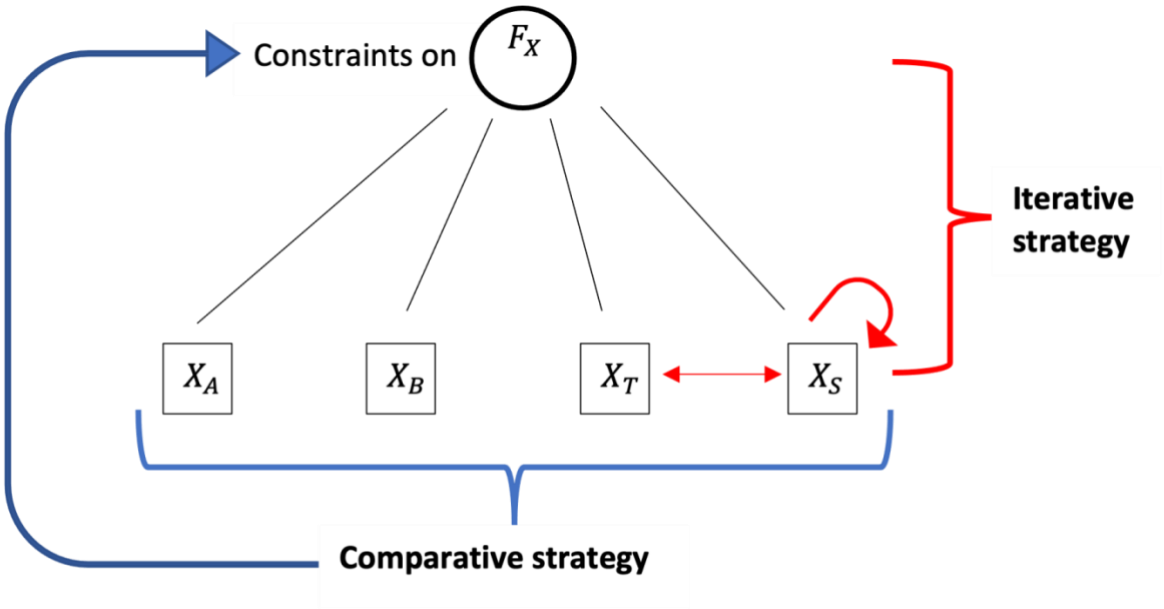

Figure 1. EMR Diagram showing how engineered multiple realization is related to the iterative and comparative strategies. F is a function that is realized by the set of realizers $\{X_A, X_B, X_T, X_S\}$, where $\{X_A, X_B, X_T\}$ are biological realizers with X_T representing the target system and X_A and X_B representing other biological systems, such as members of other species that could provide constraints on the reasoning about the relationship of the engineered model to the target . X_S is the realizer corresponding to the engineered system S, which can inform us about target system X_T or F_X through the iterative and comparative strategies respectively.

2.5 Reflections on the benefits of EMR

Framing the utility of EMR in terms of constraint-based reasoning accomplishes two things.

First, it demonstrates the limit of using multiple realization solely to support metaphysical positions like functionalism or to argue against reductionism. Consider Bechtel & Abrahamsen (2002) assessing how modellers make choices about biological plausibility,

> "[t]hose who come closest to a functionalist position are primarily concerned with developing the computational power and techniques that will best produce human-like intelligence, that is, they emphasize what result is achieved rather than how. Thus, enhancements to network designs can make them less brain-like if that is what does the job […] Other connectionists, though, lean towards the other end of the spectrum, biological realism. At a minimum, they try to keep the how from conflicting with current knowledge of the nervous system, but increasingly they seek to incorporate some of that knowledge in connectionist networks" (p.344).

On this view, those who care about neural similarity are more closely aligned with mind-brain identity theory. Thus, since cognitive models are taken to reduce to neural models, neural similarity will serve as an important factor in favour of a model. This form of biological similarity trickles into the methodology with modellers driven to include neural details in an effort to make their cognitive models more neurally plausible. On the other hand, those who care about behaviour appear to make more functionalist assumptions where the internal dynamics are not seen as significant in the modelling process. Thus, these modellers are less concerned with neural similarity from the outset.

My account of EMR complicates Bechtel & Abrahamsen's (2002) picture, allowing us to avoid the dichotomous characterization of identity theory and functionalism. On my account, it is not the case that the more reductionism-friendly modellers necessarily engage in more 'biological realism', since what counts as biological realism differs for different perspectives depending on what features they take to be biologically relevant. These modelling decisions are not tied to supporting metaphysical positions from philosophy of mind or to arbitrating disputes between philosophers of science on reductionism. Instead of a dichotomy, these modelling practices can be

interpreted as part of a spectrum where different notions of biological plausibility confer different advantages and allow us to gain different kinds of information about the role of constraints on biological and artificial systems.[17] Sometimes matching performance will count towards being more biologically realistic; other times, anatomical details will matter more to what is biologically realistic. On this view, we can resolve disputes over multiple realization by reconceiving them not as disagreement, but as reflecting different perspectives that stand open to integration.

Second, my account demonstrates philosophers of science are wrong to conclude that multiple realization has nothing to offer them. My account of EMR suggests there are avenues for the multiple realization debate hitherto underexplored, despite concerns that the debate has run its course. Connecting multiple realization to engineering shines a light on the relatively underdiscussed importance of engineering in scientific practice. My account recognises that scientists use EMRs in positive ways that further their scientific goals and support inferences that contribute to scientific progress. Not only does this better capture an important aspect of scientific practice, but my account also provides conceptual tools to understand the epistemic relation between engineering and basic science. In particular, my account proposes that engineering systems informs basic science research through constraint-based reasoning.

While the metaphysical conclusions of the MR debate have not been my focus here, I also think my account of EMR has the resources to accommodate two different views on multiple realization that currently exist in the literature. Polger & Shapiro's (2016) 'official recipe' for

[17] Compare this to Godfrey-Smith's (2008) similar picture of modelling practice: "Two scientists can use the same model to help with the same target system, while having different views about the extent and character of the similarity that the model has to the target. One might see the model as a purely predictive device. The other might see it as a causal map, a good representation of a hidden dependency structure inside the target system. And there is no dichotomy between a single realist and single instrumentalist attitude here, but a spectrum or space of possible attitudes on how model and target might be related" (p.10).

multiple realization proposes relevant similarities and differences as central for determining whether multiple realization has occurred. For Polger and Shapiro (2016), multiple realization occurs when the differences between two realizers are features that are causally relevant to the function performed. Like Polger and Shapiro's 'official recipe', my account of EMR also emphasises the importance of relevant similarities and differences by including biological similarity and relevance into what defines a perspective. However, our accounts differ in two respects. First, I don't place a causal constraint on what is considered a relevant similarity or difference. Instead, my account of EMR recognises that scientists may find other features useful for replicating functions even if they are not causally connected to the desired function. Second, while Polger and Shapiro's account is connected to their support of a modest identity theory, my account aims to avoid framing multiple realization as necessarily connected to functionalism and identity theory debates.

The main competitor to Polger and Shapiro's (2016) account of multiple realization is that of Aizawa & Gillett (2009). Their so-called 'Dimensional' view of multiple realization claims that multiple realization occurs when a property on a higher level is instantiated by two or more non-identical sets of instances on a lower level. With this as a starting point, Aizawa (2013, 2018a, 2018b) argues for three circumstances in which multiple realization is possible:

a) Multiple realization by individual differences,

b) Multiple realization by orthogonal realizers,

c) Multiple realization by compensatory differences.

These options are understood to be indicative of ways that multiple realization occurs in scientific practice. I agree wholeheartedly with the motivation of this account to capture the kinds of multiple realization that occur in scientific practice. However, the 'Dimensional' view of multiple

realization is still constrained to a fixed and static conception of what the realization relation is – a criticism also discussed in Balari & Lorenzo (2019). The benefit of EMR comes from the recognition that what counts as a realization itself will vary based on the parameters of EMR delineated above. In short, there will be some cases where a more restrictive account of realization such as Polger and Shapiro's may be conducive to a research group's goals for EMR, but there will be other cases where a more lenient account of the realization relation will be favoured. By embracing this pluralism, we allow for an account of multiple realization which is closer to scientific practice, even if it means we diverge from the origins of multiple realization's connection to reductionism and functionalism.

Finally, I believe my account can also provide a path to connect multiple realization to the ethical questions arising from it.[18] When we recognise that multiple realization is perspectival and dependent on background assumptions of scientists, it invites reflection on what values shape our concepts of multiple realization, in particular, the choice of features that determine biological relevance or similarity. For example, whether pain is multiply realized in biological systems could be useful for our scientific understanding of pain. However, it can also have social and ethical consequences in terms of our moral obligations to these organisms. The goals shaping what counts as biologically relevant for multiple realization will frequently be value-laden, whether they are scientific goals or not. The same is true when we engineer artificial systems for multiple realization. For example, concerns about the explainability of machine learning algorithms when used for automated decision-making procedures in healthcare can be informed by considering whether they are similar to human decision-making procedures. Arguments that AI systems fail to

[18] The connections between multiple realization type questions and ethical concerns is also implicit in the original work on multiple realization, see Putnam (1964).

take into account human experience into their decisions are pointing to an important kind of biological similarity that the developers of these systems fail to take into account (Chin-Yee & Upshur, 2019). This links to how we think about EMR where decisions about the set of biologically relevant constraints may be shaped by ethical concerns as much as scientific ones.

By providing a framework for considering the epistemic and ethical consequences of multiple realization in science, EMR repurposes multiple realization for the era of engineering.

2.6 Conclusion

In this chapter, I have sought to introduce a general conceptual framework for my account of engineered multiple realization. What emerges is a perspectival account of multiple realization, which specifies biologically relevant features, that determines what counts as a multiple realization in connection with researchers' goals. With this framework in place, I argue that we can rehabilitate the notion of multiple realization in philosophy of science where its main payoff is to license constraint-based reasoning about scientific phenomena based on engineered systems. This account answers calls from philosophers of science to identify the utility of multiple realization, if it is not being used to arbitrate between reductionism and functionalism. Furthermore, it provides a route to understand the utility of engineering bio-inspired computational models in cognitive science, which I turn to in the next chapter.

3.0 Applying EMR to computational neuroscience: the performance-driven and anatomy-driven approach

> *Biologists generally refer to the activity of living organisms as 'behaviour'. When talking about machines, engineers tend to use the word 'performance'. To interchange these words is to raise a smile, perhaps an appreciative smile, but the speaker risks being labelled quixotic.*
> ~ Gregory (1961), p.307

In the previous chapter, I outlined the conceptual framework of my account of Engineered Multiple Realization (EMR) and emphasized three key features: the importance of engineering in scientific practice, its explicit perspectivism, and its connection to constraint-based reasoning in science. In this chapter, I apply my account to scientific cases to demonstrate how EMR contributes to constraint-based reasoning.

3.1 Introduction

EMR, the process of engineering a system S that performs a function F also performed by a target system T, has (at least) two epistemic roles. First, it can provide greater understanding of a target system T that performs F. Second, it can provide greater understanding of the function F. In the previous chapter, I connected these epistemic roles to two corresponding strategies, an iterative strategy and a comparative strategy. However, I stopped short of applying my account to scientific cases from computational neuroscience. Here, I demonstrate how my account applies to scientific examples of network research in computational neuroscience and how this contributes to the epistemic roles of EMR.

3.2 Anatomy and performance driven approaches

I start by introducing two distinct approaches to modelling networks in computational neuroscience: anatomy-driven and performance-driven approaches. In both cases, scientists engineer systems that perform the same function in different ways. However, what they take to be different depends on their concept of biological similarity, their modelling techniques and goals. These affect what count as the norms for successful MR. On the one hand, there is an anatomy-driven approach that prioritises forms of biological similarity at the structural or cellular level. On the other hand, there is a performance-driven approach that prioritises forms of biological similarity at the cognitive level.[19]

3.2.1 Anatomy-driven approaches to computational modelling

The anatomy-driven approach to computational modelling prioritises the anatomical details of a system. Two key features of this approach are:

i) Prioritising the accuracy of anatomical features over the performance of the model,

ii) Biologically relevant features are picked in advance depending on the goals and aims of the modellers, G.

[19] These two approaches are similar to the distinction between realistic and simplifying brain models drawn by Sejnowski, Koch, & Churchland (1988). However, there are dissimilarities in that both anatomy-driven and performance-driven approaches are realistic and simplifying in different ways. While anatomy-driven models are more realistic at a particular biological level, they may simplify other parameters relevant to different scales in neuroscience. Similarly, while performance-driven models simplify by abstracting away from certain biological details, they can be more "realistic" in terms of their output and behaviour when this is an important constraint for modellers.

This is a kind of "bottom-up" modelling where the goals G shape B, what is considered relevant for biological similarity. The choice of anatomical features is then translated into formal constraints on the model depending on the modelling technique M. In other words, the modellers start from anatomical details as biologically relevant in the models to see what behaviour arises.

To see how this works in practice, I provide an example from the Chittka lab, a bee behavioural lab. The group has built simple bio-inspired computational models based on recordings of neurons in honeybee brains to investigate connections with bee behaviour. Since technological limitations prevent the imaging of the actual anatomical connections in bees, the computational models are used to give insight into what the anatomical connections could be. This approach prioritises anatomical features over performance features: "these models were not created, or indeed in any way 'tweaked' to replicate performance at any particular visual task" (Roper et al., p.3).

To do this, the researchers use empirical data on the behaviour of actual honeybees and compare this to 'simulated bees', computational models of different anatomical features in the bee brain. An important feature of honeybee neuroanatomy are mushroom bodies, collections of neurons which integrate sensory inputs, especially visual inputs in bees (Menzel, 2012). Biological systems such as the mushroom body are well-suited for computational modelling as there is a plethora of anatomical, physiological and behavioural data that can be drawn on to formalize constraints on the model (Caron & Abbott, 2017). Roper et al's two models compare potential connections from the mushroom bodies to lobular orientation-sensitive neurons providing visual input. The DISTINCT model has segregated connections from the left and right lobula neurons to the mushroom bodies such that the left mushroom body only receives input from the left lobula neuron, and the right mushroom body only receives input from the right. The MERGED model

combines the two pathways so that each mushroom body receives input from both the left and right eyes (Roper, Fernando, & Chittka, 2017, p.3).

The computational models are tested on two kinds of task: visual discrimination tasks, where the 'simulated bees' are tested on their ability to discriminate between two large patterns made up of multiple oriented bars; and visual generalization tasks, where the 'simulated bees' are trained on sets of patterns and then have to generalise from this training to novel variations of the patterns. Comparing the performance of the two models against that of real bees is meant to reveal something about how bees accomplish these tasks. For example, on the visual discrimination task, the DISTINCT model's performance was better than both the MERGED model's performance and the real bees' performance. However, when the visual stimuli of patterns were offset from the centre of the field of view, the DISTINCT model's performance accuracy dropped indicating that the DISTINCT model is not robust to these changes. In contrast, the MERGED model is not as susceptible to offset and retains similar levels of performance. The authors conclude that, while the MERGED model is not as initially accurate in the discrimination task, its robustness to cue offsetting is evidence for the MERGED model being more biologically plausible, meaning it is more likely to be the mechanism used in the bee brain.

The 'simulated bee' computational models are composed of nodes and weights, where the former approximate the lobula neurons and mushroom bodies, and the latter represent the connections between them. In this model, the weights are all set to +1 or -1 with positive weights representing excitatory connections, and negative weights representing inhibitory connections. Significantly, the modellers don't include learning so there is no updating of the weight values. This omission is indicative of a desire to abstract from particular values of weights that could optimise performance. Since the researchers' chosen biologically relevant features are the merged

and segregated connections of the simulated bees' visual system, they prioritise this over performance. They abstract from more detailed features to focus on what they have deemed most biologically relevant: the anatomical connections found in the bee brain.

Why does this count as an instance of EMR? Recall that EMR occurs when an agent engineers a system S that performs a function F in a way that is different from F's instantiation in a target system T. In this case, the simulated bees are the engineered systems S that model the function F of visual discrimination (see Figure 2 for a preview of how EMR applies to this case). The target system T is the actual honeybee. To determine whether the simulated bees (S) perform a function differently from the actual honeybees (T), the scientists rely on a conception of MR informed by a perspective dependent on their notion of biological similarity B, their modelling techniques M, and their goals G. Here, the anatomy-driven approach views anatomical features as relevant for B. So having different anatomical connections is sufficient for the function being multiply realized. This affects the choice of the modelling technique M since it means that, rather than trying to optimise performance, the modellers care specifically about the anatomical features, which are kept fixed in the model.

The purpose of such modelling is twofold and is connected to the goals G of the Chittka lab. First, it allows the Chittka lab to study how different neuronal connections affect performance, which in turn can guide hypotheses of the actual anatomical connections in the bee brain. This aligns with one use of EMRs: to provide understanding of the target system, T, in this case, the bee brain. Second, their goal is to "explore how well, or poorly, the known neuronal types within the bee brain could solve real behaviourally relevant problems and how much neuronal complexity would be required to do so" (Roper et al., 2017, p.2). The models can act as minimal models for visual discrimination tasks and identify particular features important for the functioning of the

system. This is linked to a larger goal of the Chittka lab. In their aptly-named paper, ‘Are Bigger Brains Better?’, they review research demonstrating that smaller insect brains can accomplish many complex cognitive tasks with fewer neurons than mammal brains (Chittka & Niven, 2009). This is linked to a second use of EMRs: to provide understanding of the function, F, in this case, visual discrimination.

3.2.2 Performance-driven approaches to computational modelling

The performance-driven approach primarily seeks to optimise performance before turning to the role of biological similarity in a model. The key features of this approach differ from the anatomy-driven approach in the following ways:

i) Prioritising performance similarity rather than anatomical similarity imposed by modelling a biological system. In doing so, this allows different possible structural organisations of the system.

ii) Relevant structural features are identified only after achieving the desired performance threshold. These biological features may depend on the tools available to the researchers, which affect the forms of biological similarity that can be assessed.

In the performance-driven approach, the choice of the modelling methodology M is primary, I consider the use of Deep Neural Networks (DNNs). This in turn affects how the scientists conceive of B, what is biologically similar in the model. Initially the choice of biological similarity is based on performance rather than anatomical features. However, once the performance threshold is achieved, the modellers can look for more similarities at the structural or anatomical level.

The use of DNNs as models for the brain in neuroscience research is a perfect example of the performance-driven approach that is fast becoming popular in contemporary computational neuroscience. One significant contemporary example is found in Yamins & Dicarlo (2016) who characterise their method as "performance-based optimisation, in which the parameters of large multi-layer neural networks are chosen to optimise the networks' performance on a high-level ecologically valid visual task" (p.114). In basic terms, an artificial neural network with multiple layers of nodes will rely on a learning rule (often back-propagation) to update the weights between nodes. Many of these neural nets are trained through supervised learning where they are trained with labelled data and scientists already know the desired output (e.g., images of objects that have already been labelled by humans). Then a cost function sets how the network matches up to the desired target and updates the weights. The aim is to get the best possible set of weights to enable the network to make the correct mapping from inputs to the desired outputs for untrained images. For example, a network may be trained on labelled images of birds in order to learn how to correctly distinguish photos of an owl from those of a flamingo. Since much of this research has been developed for commercial or military purposes, an important question is: what does this have to do with neuroscience?

The neuroscientific motivation underlying these engineered systems can be found in Fukushima's (1980) development of the Neocognitron, an early predecessor of today's convolutional neural networks.[20] He opens his paper noting: "it seems to be almost impossible to reveal [the mechanism of pattern recognition in the brain] only by conventional physiological experiments. So, we take a slightly different approach to this problem. If we could make a neural

[20] Though the sentiment at least extends as far as early cybernetics research with Rosenbleuth & Wiener (1945) also describing the utility of models in similar terms.

network model which has the same capability for pattern recognition as a human being, it would give us a powerful clue to the understanding of the neural mechanism in the brain" (p.193). This same goal is echoed almost 40 years later by Yamins & DiCarlo (2016): "selecting biologically-plausible neural networks for high performance on an ecologically-relevant sensory task will yield a detailed model of the actual cortical areas that underlie that task" (p.117). The idea is that building something that can perform as well as the brain will be useful for understanding the brain's internal dynamics, allowing so-called 'synthetic neurophysiologists' to study and experiment on a system *in silico* (Kriegeskorte, 2015, p.16). This is one of the main goals of the performance-driven approach: to create an artificial system that can serve as a proxy for a biological system (Tarr & Aminoff, 2016). More generally, the EMR acts as a proxy for the target system T.

There are some important differences to highlight between the anatomy-driven and performance-driven approach based on how they conceive of biological similarity. They clearly differ about what they initially take to be biologically relevant since the performance-driven approach is committed to performance features and the anatomy-driven approach focusing on anatomical features. However, since the model is meant to act as a proxy for a target system in the performance-driven approach, once you have a model that accurately captures performance, the idea is that scientists can use it to investigate other biologically relevant features. This is described succinctly in Kriegeskorte's (2015) manifesto for the performance-driven approach: "The challenge ahead is, first, to scale recurrent neural net models for vision to real-world tasks and human performance levels and, second, to fit and compare their representational dynamics to biological brains" (p.15). Computational neuroscientists start by building a model that best approximates human performance and then comparing this against neural dynamics.

Why does this count as an instance of EMR? In this case, the DNNs are the engineered systems S that model the function F of object recognition (see Figure 3 for a preview of how EMR applies in this case). The target system T is the human visual ventral stream. To determine whether the DNNs (S) perform a function differently from humans (T), the performance-driven approach views performance features as relevant for B. This is because here biological similarity is being used to establish that the same function is being performed. Once the threshold level of performance is established, the DNNs can then be used to investigate other forms of biological similarity B to decide whether these count as multiply realized or not.

This process means that, unlike the anatomy-driven approach, where the biological features are fixed in advance, in a performance-driven approach, the biologically relevant features can expand after reaching a threshold level of performance. What is selected as biologically relevant at this point is interest-dependent. Some researchers look to neural representations to assess biological similarity by examining the networks' ability to predict neural activity in layers of the ventral visual stream (Kriegeskorte, Mur, & Bandettini, 2008; Schrimpf et al., 2018) Other researchers instead look to behavioural similarity investigating whether neural networks rely on the same features as humans when categorising objects (Geirhos et al., 2019). Due to this fluidity in what counts as biologically relevant, even within the performance-driven approach, you may still get different norms dictating what counts as successful EMR.

3.3 Computational modelling according to EMR

In this section, I explain how these two examples support constraint-based reasoning through the iterative and comparative strategies. As a perspectival account of multiple realization,

the goals and payoffs of EMR for a scientist will vary widely with no one overarching motivation but could include the following: the ability to perform artificial experiments in silico, an existence proof that artificial systems can replicate human performance, the creation of minimal models of particular functions and cognitive abilities, improving performance of computational models, and constraining the space of possibilities for mechanisms underlying cognitive processes. These goals suggest that the generation of EMRs is in line with non-reductionist strategies in science leaving space to ponder its relation to functionalism.[21] However, for the purposes of this chapter, I am more interested in establishing a positive role for multiple realization that is not tied to support of functionalism.

To gain insight into why scientists would generate EMRs, recall EMR's two epistemic roles. The process of engineering a system S that performs F could:

1) Provide greater understanding[22] of target system T (that also performs F),
2) Provide greater understanding of reproduced function F.

Each of these uses of EMR correspond to a particular strategy that supports constraint-based reasoning. First, there is an iterative strategy linked to greater understanding of target system T, where the regulative ideal is to eliminate MR. Second, there is a comparative strategy linked to greater understanding of the reproduced function F, where the regulative ideal is to maximise MR.

Here, I will discuss some tools utilised in EMR research and their associated scientific goals. There are three ways that scientists work with constraints in the EMR examples I provided:

[21] Sober (2010), for example, gives a non-metaphysical reading of functionalism as "an empirical thesis about the degree to which the psychological characteristics of a system constrain the system's physical realization" (p.227). This reading seems compatible with my constraint focused account of EMR.

[22] There is a large literature on understanding in the philosophical literature. But for my purposes, I am using it loosely here rather than connecting it to any particular account of scientific understanding.

a) *Constraint mapping* – the constraints on the model undergo an iterative process to better align them to the constraints on a target system T (whether law-like, causal, historical or otherwise),
b) *Constraint identification* – constraints on the model help identify potential constraints on the class of systems realizing F,
c) *Constraint refinement* – constraints are refined into the minimum number of constraints required to generate some pre-determined outcome (i.e., a minimal model).

Each of these can be associated with a particular goal. Constraint mapping takes place in order to create systems that act as artificial proxies for experimentation in silico. Here EMR is tied to generating more information that maps on to the target system, T, so scientists can understand it better. Constraint identification is linked to exploration where new questions or potential hypotheses about a system are discovered. Here EMR can tell us something about the target system T or the general function F that a target system instantiates. Constraint refinement is linked to the generation of minimal models[23], where the goal is to better understand the system by identifying a minimal set of constraints sufficient for an outcome. Here EMR is focused on providing more understanding of the general function F. I relate these tools to the goals of the anatomy and performance-driven approaches.

[23] There are at least three different notions of minimal models in the literature. Chirimuuta (2014) draws a helpful distinction between A-minimal models, which incorporate key causal factors into a model to see the behaviour they generate (Weisberg, 2012), B-minimal models, which are used to explain macroscopic patterns of behaviour in heterogeneous systems (Batterman & Rice, 2014) and I-minimal models, which describe the information processing capacity of neurons and abstract away from biophysical details (Chirimuuta, 2014). While EMRs could be connected with different sorts of minimal models, here the Chittka lab research seems like a mix between an A-minimal model and I-minimal model.

3.3.1 Anatomy-driven approach

First, I will use the anatomy-driven approach to illustrate how the comparative strategy is used to learn about potential constraints on a function. The comparative approach aims to maximise MR in order to contrast different systems that realize F in order to gain understanding of the constraints on F. This is similar to the process discussed in Green et al. (2018), whereby systematic comparison of different species and the variations between different organisms helps guide the search for generalisations and design principles. In other words, this can inform scientists about potential law-like (L) and enabling (E) constraints. In these cases, computational modellers will engineer systems that perform a function in several different ways along the dimension(s) of interest.

The case I discuss from the Chittka lab serves as a very simple example of the comparative strategy. Here the simulated bees that vary with respect to their anatomical connections act as two rival realizers of a function compared to the biological target system T. This is pictured in Figure 2. Scientists use both constraint identification and constraint refinement methods. Here a function F is specified by thinking of biological similarity B as anatomical features, i.e.) whether neurons receive merged or separate inputs. This is then translated into a formalised constraint in the model determined by the connections between different nodes. Since anatomical similarity is most important, building simple neural networks constrained by anatomical data is ideal for their scientific investigations. By comparing the DISTINCT and MERGED simulated bees, the scientists can select the more robust performance as indicative that the anatomical connections in the MERGED model are more successful for the function F (here, visual discrimination). This process is a form of constraint identification, where the constraints on the model inform both the understanding of the function F and the target system T. First of all, it tells the researchers that

merged inputs could be generally advantageous for realizing visual discrimination. This provides information about a potential anatomical constraint on the performance of the function of visual discrimination, which can in turn lead to the development of new hypotheses. Second, the model informs scientists about constraints on the anatomy of the bee neuronal system. This is useful because the anatomical constraints cannot be identified in the biological system since we do not yet have the technology to observe and intervene on Kenyon cells in the bee brain. In this case, robustness to perturbation of orientation serves as a guide to what the bee brain could look like. The researchers are also involved in the process of constraint refinement. This is because the simulated bees serve a role in a broader project of finding minimal models of complex behaviour such as visual discrimination tasks.[24]

One concern about the comparative strategy is whether it can discover universal generalizing principles that apply more broadly to all systems. This concern may be particularly pertinent for DNNs which have important differences with biological systems or are subject to concerns about overfitting. However, the comparative approach can still be useful even if it does not eventually lead to general principles. As I have discussed, the comparative strategy can also inform us about the target system T and later I will explain how it can be used in tandem with the iterative strategy.

[24] Interestingly, the goal in this case parallels what Koskinen (2019) describes in synthetic biology work that tries to engineer minimal genetic systems. This provides further evidence that EMR is a more widespread phenomenon than is currently allowed for in the MR literature and may suggest there are similar features and goals between synthetic biology and computational modelling of cognition that prompt the adoption of such techniques.

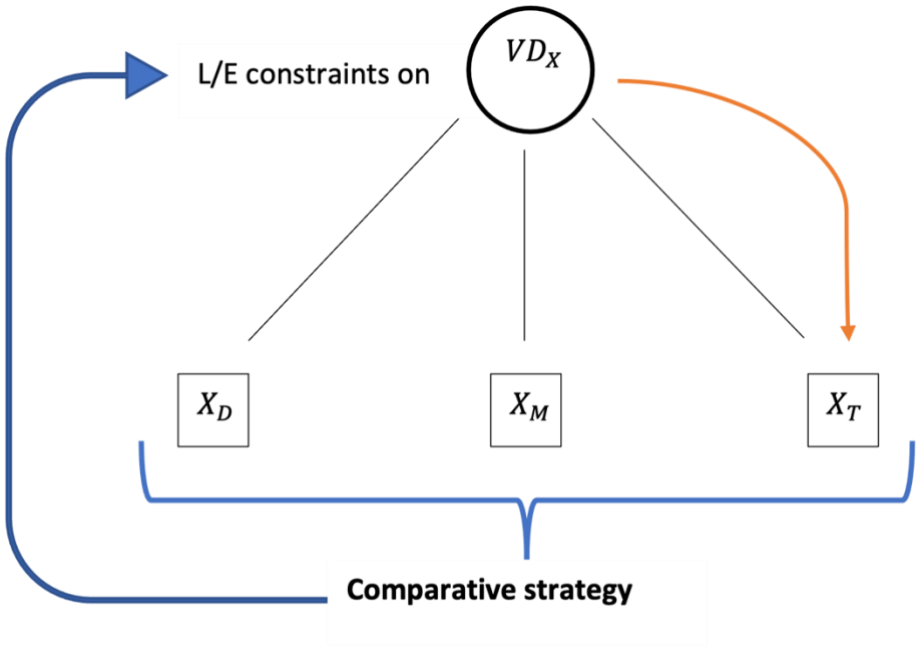

Figure 2. Anatomy-driven EMR Diagram showing how EMR is used in the anatomy-driven approach example from the Chittka lab. VD_X is the function, visual discrimination, that is realized by the set of realizers $\{X_D, X_M, X_T\}$. X_T is the target system – in this case, an actual bee. X_D and X_M are the engineered realizers – in this case, the simulated bees, the DISTINCT and MERGED network respectively. The comparative strategy leads to constraint identification of law-like and enabling constraints, which subsequently inform ideas about the target system, X_T.

3.3.2 Performance-driven approach

Second, I will use the performance-driven approach to demonstrate how constraint mapping and the iterative strategy work. In this research, scientists are engaged in a process of constraint mapping. The primary constraints are performance constraints, specifically meeting a threshold for successful object categorization. These selected constraints cannot be entirely separated from the choice of the modelling technique. It is because DNNs have had success in object recognition that they could then become viable candidates for a model of the visual ventral

stream. Unlike the anatomy-driven approach, here replicating performance-level details is important for sameness of function and successful EMR. However, this is in tension with one of the broader purported goals of the performance-driven approach, which is to create models that allow for *in silico* experiments to further understanding of the human brain. On this goal, neural similarity, rather than just performance similarity, is desired and it is impossible to determine this on the basis of their performance similarity alone.

To achieve this goal, constraint mapping is used. In constraint mapping, researchers use an iterative strategy to bring biological and artificial realisers more in alignment, thereby trying to eliminate MR along the dimension(s) of interest. This is shown in Figure 3. It is a process of selecting the notion of biological similarity B to fit the researchers' goals G by using more fine-grained constraints on the model. For example, when humans do object recognition, they have a shape bias, meaning they will primarily use the shape of objects to classify them. Typically, DNNs have been interpreted as exhibiting this shape bias too (Kubilius et al., 2016; Ritter et al., 2017). However, this interpretation of DNNs has also been challenged. First of all, neural networks are often susceptible to "adversarial examples", images that trick networks into making incorrect categorizations. The issue for neural networks is that many of these adversarial examples are ones that humans are supposedly not susceptible to. For example, adding random noise unnoticeable to the human eye to input images can be enough to lead a neural network to wrongly categorise an object (Goodfellow et al., 2014; Kurakin et al., 2016). Buckner (2018) questions whether adversarial examples always threaten the utility of deep convolutional neural networks. However, I contend these adversarial examples are an issue for the use of the model as a proxy for the biological system, since the similarity of the features used for categorisation is a relevant criterion

to use in assessing whether we can use DNNs as models of the human visual system (Lopez-Rubio, 2018).

Some adversarial examples show that, rather than demonstrating a shape bias, DNNs are more sensitive to visual texture - the characteristics of the surface of an object such as the crisscross visual texture of birds' feathers (Geirhos et al., 2019). This texture sensitivity means, while the neural network is a successful EMR with respect to some aspects of performance that are of interest to scientists, it will not be successful in others, such as in aiding the goal of performing *in silico* experiments on an artificial brain-like system. Rather than give up on the model, researchers instead try to rectify this by adjusting the performance to more closely match human performance (Geirhos et al., 2019; Peterson et al., 2018). For example, in order to induce a shape bias, Geirhos et al. (2018) train the network on a new dataset that forces the network to rely on more than texture. They find that inducing a shape bias in this way also leads to greater robustness in dealing with distortions. Here constraint mapping leads to improvement of performance and progress towards scientists' goals. To help them quantify their progress towards a more biological model, computational neuroscientists use benchmark challenges such as the Brain-Score platform, which consists of two neural benchmarks, where ANNs are compared to neural recordings from areas V4 and IT in macaque monkeys, and one behavioural benchmark, where ANNs are compared to behavioural data from humans. This serves to test whether DNNs with better performance will be better models of the primate visual ventral stream (Schrimpf et al., 2018).

However, it is unlikely that this iterative approach can ever address all the differences between DNNs and the brain so it will not be possible to completely eliminate multiple realization in these engineered cases. This is because, as EMR is a perspectival account of multiple realization, there are multiple notions of multiple realization at play depending on what you construe as

biologically relevant given your goals and techniques. For example, a neuroscientist who is interested in focusing on the brain purely in terms of information processing may not view substrate as biologically relevant for multiple realization. But for a scientist who does take this to be a relevant difference, difference in substrate means there would be multiple realization. So we can see that, even on an iterative strategy, when you minimise or eliminate some forms of multiple realization, other scientists may still view the same system as a multiple realizing a function.

Furthermore, the choice of how to model a system already restricts what can be captured in the model. For example, DNNs are unlikely to ever adequately address the role of embodiment without being coupled with a robotic 'body'. Thinking we can eliminate all multiple realization quickly descends into the view that only a perfect representational model is sufficient; as Rosenbleuth and Wiener (1945) observe "the best material model for a cat is another, or preferably the same cat" (p.320). However, they note that the models used in science are always likely to involve abstractions and will necessarily be partial. The same point must be heeded for multiple realization. For this reason, we can view the elimination of MR as a regulative ideal, which guides the iterative strategy but cannot be fully achieved. The hope is that as more biological constraints are introduced into a model, the less likely it is that multiple realization occurs.

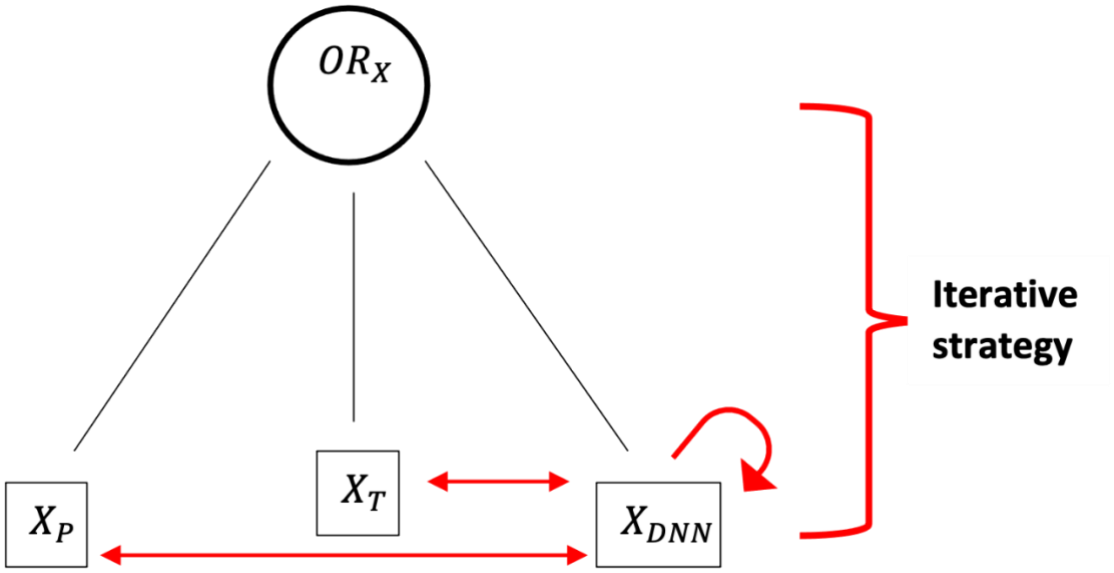

Figure 3. Performance-driven EMR. Diagram showing how EMR is used in the performance-driven approach examples with DNNs. OR_X is the function, object recognition, that is realized by the set of realizers $\{X_{DNN}, X_P, X_T\}$. X_T is the target system – in this case, the human visual ventral stream. X_P is another biological realizer – in this case, the non-human primate visual system. X_{DNN} is the engineered realizer – in this case, the Deep Neural Network. The iterative strategy leads to constraint mapping between the DNN and human realizer and/or between the DNN and non-human primate realizer.

3.3.3 Combining the two approaches

In summary, I have explained how EMRs developed in the anatomy-driven and performance-driven approaches contribute to constraint-based reasoning through the comparative and iterative strategies.

Here I want to emphasise that the anatomy-driven approach should not be viewed as homogenous. There will be anatomy-driven approaches that generate EMRs in order to use for constraint mapping and the iterative strategy as well. In these cases, constraint mapping will be used to make closer connections with a biological target system. As we saw, the Chittka lab adopts

a high level of abstraction in their models by setting weights to +1 and -1 for excitatory and inhibitory connections respectively. However, one could select weights that more closely correspond to an actual biological system. Delahunt et al. (2018) use a form of constraint mapping when they set input parameters in their model from *in vivo* data based on measurements from moths. This reflects a difference in goals. The Chittka lab are generating minimal models that can identify sets of features sufficient for performance. In contrast, Delahunt et al.'s model can make more precise inferences about a specific target organism. By reducing MR through constraint mapping, the latter aim to gain more understanding of the particular target system rather than of the function F.

Similarly, the performance-driven approach is not homogeneous and can be used with a comparative strategy. For some, this is an important benefit of DNNs because they allow one to generate a large number of different realizers of a particular behaviour: "one can easily train thousands of networks while systematically varying the behavioral task, neural architecture and cost functions, thus enabling the study of large ensembles of potentially different solutions to a given behavioural task" (Musall et al., 2019, p.234). Other research shows how DNNs can be used in constraint identification. Kell and McDermott (2019) argue for the value of DNNs as providing insight into task constraints, where the similarity between DNNs and the human ventral stream is due to constraints that arise from biological vision tasks. On my framework, this suggests the potential of DNNs to inform us about tasks as enabling constraints (E). Constraint identification can also end up informing scientists about biological target systems as well. For example, Haesemeyer et al. (2019) train artificial neural networks on navigating heat gradients to better understand zebrafish behaviour. They use an explicit performance-driven approach: "[we avoided] constraining network representations by anatomy, but instead limit[ed] constraints as much as

possible to the behavioral goal and the available motor repertoire" (p.1125). The authors train two models: one trained to predict the consequences of movements in the heat gradient, one trained using reinforcement learning for successful behaviour. Having discovered similarities between their trained models and zebrafish neural networks, they performed lesioning experiments comparing random removal of units to removal of specific units that were deemed most similar to zebrafish responses. Removing the latter significantly decreased performance in the gradient navigation tasks. They conclude that the more 'fish-like' units are important for gradient navigation. In doing so, the use of DNNs leads to hypotheses identifying constraints.

Finally, the comparative and iterative strategies are not mutually exclusive. The comparative strategy can be used in tandem with the iterative strategy and vice versa. The comparative strategy can serve to identify relevant constraints, which are then fed into an iterative strategy to develop computational models that are more biologically plausible. For example, in connectome research, Izquierdo & Beer (2013) generate an ensemble of possible models of neural circuits (not DNNs), which in turn helps formulate experiments to distinguish between various possibilities. In this case, the process continues with these experiments finding further constraints that are added to subsequent modelling.

The examples presented in this section present cases of different approaches using various strategies or sometimes the same strategies in tandem in order to complicate the picture. The performance-driven and anatomy-driven approaches are not meant to create a new dichotomy to replace functionalism and reductionism (see §2.5 for further thoughts on this). These two approaches do not constitute a dichotomy but rather a flexible spectrum, where each may use the tools available through EMR for their own goals.

3.4 Conclusion

In this chapter, I sought to demonstrate how my account of EMR can be applied to scientific cases to illuminate scientific practice. I introduced two approaches to using networks in computational neuroscience: an anatomy-driven approach and a performance-driven approach. These examples help demonstrate the three key features of my account of EMR. First, the cases I consider in this chapter are examples of neuroscientists engineering systems to reproduce particular functions. Second, by showing that the anatomy-driven and performance driven approaches introduce different forms of multiple realization, I have demonstrated the utility of a perspectival account of multiple realization. Third, I have explained how these cases of EMR contribute to constraint-based reasoning, thus showing the utility of multiple realization to philosophy of science.

4.0 Engineering artificial model organisms: a role for Deep Neural Networks in neuroscience?

> *A connectionist unit – of whatever type – is a mythical beast, as elusive in the biological world as the gryphon.*
> ~ Boden (2008), p.1115

There has been an explosion of interest within computational neuroscience in the use of deep neural networks as models of the brain. But what kind of model is a deep neural network? In this paper, I defend the view that deep neural networks share similarities to model organisms and 'Krogh organisms'. Then, using my EMR (Engineered Multiple Realization) framework, I consider the implications for what inferences can be drawn from them while discussing two accompanying methodological strategies in computational neuroscience research. In conclusion, this identifies a role for deep neural networks in neuroscience and a motivation for why scientists attempt to engineer artificial model organisms.

4.1 Introduction

Recently, there has been marked enthusiasm surrounding the use of deep neural networks (DNNs) in neuroscience (Kriegeskorte, 2015; Marblestone, Wayne, Kording, & Scholte, 2016; Richards et al., 2019; Yamins & DiCarlo, 2016). However, since DNNs are only loosely inspired by biological neural networks and mostly engineered for purposes other than understanding the brain, this raises questions of what their role is within neuroscience. Specifically, in what sense

are they a model of the brain? While this topic is little discussed in the philosophical literature,[25] this question has prompted several proposals from computational neuroscientists categorising DNNs as ideal observer models (Kell & McDermott, 2019), exploratory models (Cichy & Kaiser, 2019), direct-fit models (Hasson et al., 2020), mechanistic models (Lindsay, 2020) and artificial model organisms (Musall et al., 2019; Scholte, 2018). In this paper, I defend the view that DNNs are most similar to artificial 'Krogh organisms', a type of organism introduced in Green et al. (2018), though DNN researchers are striving to use them as model organisms. Research organisms can be used to support constraint-based inferences in different ways, which is reflected in two sets of methodological strategies used in DNN research. Examining the kinds of inferences made using these strategies elucidates the role of DNNs and neuroscientists' attempts to engineer artificial organisms in the first place.

In section 2, I discuss Ankeny and Leonelli's (2012, 2020) account of model organisms and Green et al.'s (2018) account of Krogh organisms and contrast them. In section 3, I apply these accounts to DNNs and argue that DNNs exhibit some features of both. They are sometimes striving for model organism status but are often more accurately interpreted as Krogh organisms. In section 4, I use my account of engineered multiple realization (EMR) to frame different uses of DNNs and the methodological strategies that accompany them. With this in mind, I claim that, depending on their use in an iterative or comparative strategy, it is appropriate to use DNNs as model organisms or Krogh organisms respectively. In section 5, I respond to other accounts of DNN modelling in the literature and consider some objections to my account. I conclude that my account offers a role

[25]Stinson (2020), a rich analysis of connectionist models through the lens of the models and modelling and explanation literature, is a notable exception.

for DNNs within computational neuroscience as well as identifying strategies for how scientists use them.

4.2 Model organisms vs Krogh organisms

Animal models are a crucial part of scientific practice in the biological sciences. However, animal models are often used in very different ways, reflecting their varied roles in scientific research. In this section, I offer a brief characterization of philosophical accounts of the use of two kinds of animal model: model organisms and Krogh organisms. Though both of these use animals as models, they differ in terms of their research practices and goals. Here I explain the differences between them in order to later apply them to DNNs.

4.2.1 Model organisms

Model organisms have long been a significant focus of scientific investigation in the biological sciences. They have become a prevalent feature of research in neuroscience culminating in well-developed research programs, particularly using rats and mice (Yartsev, 2017). In line with their ubiquity, there is a correspondingly large historical and philosophical literature on model organisms (Ankeny & Leonelli, 2011; Bolker, 2009; Dietrich et al., 2019; Kirk, 2012; Kohler, 1994; Levy & Currie, 2015; Meunier, 2012; Milani & Ghiselli, 2019; Rader, 2004). At first pass, model organisms can be defined as non-human species studied in detail to develop understanding and theories of biological phenomena that can then be applied to other organisms (Ankeny & Leonelli, 2011, p.2). However, this does not serve to demarcate model organisms from other uses

of animal models in the biological sciences. I focus on Ankeny and Leonelli's account of model organisms, which contends that model organisms have specific attributes such that they constitute a unique kind of scientific practice.

Characteristic examples of model organisms are those officially recognised by the American NIH (National Institute Health) as used for biomedical research. These include non-human mammalian model organisms such as mice, as well as non-mammalian model organisms such as fruit flies, zebrafish, and budding yeast (Ankeny & Leonelli, 2020). It is generally recognised that model organism research typically involves forms of standardisation that rely both on material features of the organism and social features of the research program (Ankeny & Leonelli, 2011; Meunier, 2012). Material features are the characteristics of a model organism that make it conducive to standardisation. For example, genetic standardisation requires an organism with short life cycles, high fertility rates and susceptibility to genetic modification techniques. For this reason, mice and fruit flies fit the kind of organism that fulfils the criteria for being a model organism. Additionally, to achieve standardisation, social and structural features are needed to generate an infrastructure supporting this research process. For example, shared community resources such as databases like Flybase (Larkin et al., 2021) or stock centres allow different researchers to have access to the same standardised organism to support generalisations. This consequently affects the community norms within model organism research where strong social ties and research sharing practices are encouraged.

On Ankeny & Leonelli's (2011, 2013, 2020) account, these material and social features must be connected to the epistemological goals of model organism research communities, which can be understood through two key concepts:

a) Representational scope - “how extensively the results of research with a particular experimental organism (a specimen or token) can be projected onto a wider group of organisms (a type)” (2011, p.7),
b) Representational target - “the phenomena to be explored through the use of the experimental organism” (2011, p.8).

Ankeny and Leonelli take the key features of model organisms to be their wide representational scope and wide representational target.

First, model organism research relies on claims of representational scope to justify the value of model organisms in generalising to other organisms. While non-model experimental organisms can be investigated as a means to gain greater understanding of an individual organism, model organisms are studied so that the results can be projected on to other organisms. For example, if a mouse is being used as a model organism for clinical research on Parkinson’s, the motivation is to use this research and project the results on to other organisms, such as humans. Thus, a model organism should have a wide representational scope to allow for results obtained from model organism research to be generalised to other organisms.

Second, model organisms are established to encourage integrative research practices. Instead of being solely used to study one particular phenomenon, model organisms are used for more holistic understanding that will integrate knowledge about the organism “in terms of their genetics, development, and physiology, and in the longer run, of evolution and ecology, among other processes” (Ankeny & Leonelli, 2013, p.24). For example, model organism research communities will collect various sets of data on the organism so that it can be used as a standard organism for lots of different kinds of research. A mouse can be used as a model organism in Parkinson’s research as well as for other clinical or cognitive research. Thus, a model organism

will have a broad representational target encouraging its application to researching a wide range of phenomena.

4.2.2 Krogh organisms

Another form of animal model research is the use of 'Krogh organisms'. The name derives from physiologist August Krogh's principle: "For a large number of problems there will be some animal of choice or a few such animals on which it can be most conveniently studied" (Krogh, 1929, p.247). This reflects the fact that the choice of a research organism in these cases is based on trying to solve a particular problem rather than using a standard model organism. 'Krogh organisms' are organisms that display specialised, often extreme, adaptational features making them ideal for investigating a specific biological problem involving that adaptation (Green et al., 2018). For example, marine organisms such as the Chinese mitten crab (*Eriocheir sinensis*) are used to investigate osmoregulation. Here the Chinese mitten crab has had to adapt to living in high-salinity environments, making it a useful guide for understanding the potential mechanisms and genes underlying osmoregulation (National Research Council, 1999; Yang et al., 2019). The adaptive traits of Krogh organisms can often also be useful in making particular mechanisms easier to investigate. Green et al. (2018) interpret Krogh's principle as a heuristic where Krogh organisms act as "experimental access points" to investigate a particular mechanism or phenomenon (p.8). For example, in neuroscience, Hodgkin and Huxley used *Loligo* squid as their Krogh organism to measure action potentials since the large size of the giant axon made it well-suited to recording action potentials through voltage clamping.

Krogh organisms differ from model organisms with respect to both representational scope and target. First, their representational scope is not established meaning their generalisability is

open to empirical investigation. While model organisms rely on being generalisable, Krogh organisms do not need to be in order to be helpful for understanding a particular phenomenon. Because of this, whether or how they generalise to other organisms is a matter for empirical investigation. To do this, a comparative approach is used where a Krogh organism must be compared to other organisms with different characteristics or organisms subject to different developmental constraints. This comparative approach allows researchers to gather information on generalised principles relating the structure, functions, and constraints on organisms.

Second, Krogh organisms have a narrow representational target. Krogh organisms are problem-focused meaning their utility is based on their convenience in understanding a specific problem or phenomenon, such as osmoregulation. By studying an organism with an extreme adaptation, such as the Chinese mitten crab, the hope is that this will shed light on the relationship between structure and a particular function that would be much harder to detect or manipulate experimentally in organisms without the extreme adaptation. Then, the structure-function relation found in the Krogh organism may in turn apply to other organisms without extreme adaptations but that face similar problems. Axolotls, with their ability to regenerate limbs, may serve as Krogh organisms to help inform research on limb and spinal cord regeneration (Russell et al., 2017). To do so, Krogh organisms do not need to be similar to target organisms. On the contrary, they often act as 'negative models' selected because of their dissimilarity to other species and can highlight the shortcomings of generalising knowledge derived from more established model organisms.

As is apparent, the difference in representational scope and target for model and Krogh organisms leads to methodological differences in their use for research. While model organisms are accompanied by a well-developed scientific infrastructure in order to support standardisation,

Krogh organisms are more typically unusual experimental organisms with less focus on the integration of different forms of information about them.

4.3 DNNs as Krogh organisms

In this section, I use the accounts of model and Krogh organisms outlined in the previous section to assess whether DNNs can count as artificial versions of these organisms. While machine learning can be used in several different ways in neuroscience such as for identifying predictive variables, I shall be mainly interested in the use of DNNs as computational models of the brain (Glaser et al., 2019). In considering how to interpret DNNs as models, there have been recent proposals by scientists relating them to model organisms.[26] Scholte (2018) first likened DNNs to model organisms claiming that computational modelling provides a menagerie of 'DNimals' to experiment on, serving a similar function to animal model organisms. Shortly after, Musall et al. (2019) coined the term 'artificial model organism' to describe the use of DNNs as an approach to connecting neural dynamics to behaviour. On Ankeny and Leonelli's account, the central features of model organisms are standardisation as well as the material and social features contributing to a model organism's representational target and scope. How do these apply to DNNs? I will consider these features in the next subsections.

[26] See also Jonas & Kording (2017) who do not explicitly argue for DNNs as model organisms but analyse microprocessors as "engineered model organisms" (p.2).

4.3.1 Standardisation

With respect to standardisation, much of Ankeny and Leonelli's (2011, 2013) account centres on genetic standardisation but a genetically-based approach is not necessary for a conception of model organisms and their use. Leonelli et al. (2014) suggest different forms of standardisation with parallels to DNNs: behavioural and environmental standardisation. They examine the use of animal models in alcoholism research suggesting that the environment and experimental set-up become part of the standardisation process in order to increase the validity of the animal model's use in modelling the human phenomenon. This has apparent analogues to DNN research, where these forms of standardisation are achieved through training procedures, datasets and task selection. In terms of training procedure, many DNNs are trained through supervised learning, where the target output is known in advance, or by reinforcement learning, where networks are rewarded for taking certain actions. In both cases, this allows researchers to exert some control over the target behaviour of the network, contributing to standardization.

The choice of tasks has a dual purpose of serving as standard benchmarks for comparison and to reinforce ecological validity. Tasks themselves have been described as model organisms by the computer science community, for example, chess was historically described as the *Drosophila* of AI owing to remarks by Russian mathematician Alexander Kronrod in 1965 (Ensmenger, 2012; McCarthy, 1990). While this rhetorical flourish certainly served to advertise chess as a productive research program for progress in artificial intelligence,[27] it also reflects the role tasks play for standardisation in computational research. Just as *Drosophila* researchers standardised their

[27] See Ensmenger (2012) for more context on the way chess shaped AI research and its limitations compared to Drosophila research.

methods and procedures over time to make their results more translatable to other contexts, standardising tasks in AI allows for easier comparisons to be made between networks.

Aside from standardisation, tasks are also used to reinforce ecological validity when inspiration is drawn from animal research to find tasks for AI. Due to the no-free-lunch theorems, there is no general-purpose learning algorithm that will be able to perform well on all problems so researchers must pick the relevant set of problems to focus on (Wolpert & Macready, 1997). In deep learning, this led to the selection of the 'AI-set', which is the set of tasks that is meant to contain "all the tasks involved in intelligent behavior" including visual perception, planning and control (Bengio and LeCun 2007, p.4). These are selected based on what "most animals can perform effortlessly" and "tasks that higher animals and humans can do" (Bengio & LeCun, 2007). Richards et al. (2019) note that this set of tasks dovetails with the focus of neuroscientists on "the behaviors or tasks that a species evolved to perform" (p.1763). The importance of using ecologically valid tasks is often alluded to, such as in Yamins and DiCarlo (2016): "selecting biologically-plausible neural networks for high performance on an ecologically-relevant sensory task will yield a detailed model of the actual cortical areas that underlie that task" (p.117). Hence task selection shores up ecological validity, which in turn justifies the utility of DNNs for understanding human and animal cognition.

The features of standardisation and ecological validity are particularly clear in examples of AI challenges, such as the 'Animal-AI Olympics' (Beyret et al., 2019; Crosby, Beyret, & Halina, 2019). The Animal-AI Olympics is a recent AI competition that translates vision-based animal cognition tasks into a virtual environment to test the cognitive capacities of AI. The participants do not know the tasks in advance and are told to submit AIs that they believe will display robust food retrieval behaviour. The tests range from simple food retrieval to tasks that require object

manipulation in the virtual environment. First, in the Animal-AI Olympics, the tasks serve both as a standardising benchmark to compare AI competitors as well as a way to establish comparisons between networks and animals. In this respect, the development of benchmarks can scaffold a comparative strategy of the sort used with Krogh organisms.[28] Second, the Animal-AI Olympics uses these tasks as a way to reinforce ecological validity. Crosby (2020) argues that, unlike older AI challenges such as computer chess, the Animal-AI Olympics provides us with a test that is relevant to an understanding of cognition that is based on sensory and perceptual abilities. The ecological validity and relevance come from adopting and developing tasks based on what is typically used to test animals.

In animal model organisms, standardisation is enabled and enhanced through material and social features. Does the same hold for our candidate artificial model organisms? Model organism research was often supported by the sharing of information through newsletters such as the 'Drosophila Information Service' starting in 1934 (Kelty, 2012). Similarly, researchers utilize shared resources such as Flybase, a repository of genome data for Drosophilidae. This has parallels to the establishment of shared databases in DNN research such as ImageNet, which is a large image dataset used to train neural networks, and shared simulation platforms and software libraries, such as TensorFlow and PyTorch.[29] For *Drosophila* research, social infrastructure was also developed through establishing an annual Drosophila Research Conference beginning in 1959. The beginnings of this infrastructure can be seen in DNN research with the creation of an annual conference called 'Cognitive Computational Neuroscience', which connects communities in

[28] Though see Firestone (2020) for concerns about whether such test sets really provide a successful benchmark for comparison.

[29] Though, given the issues reported with these large image datasets as discussed in Prabhu and Birhane (2021), one might question the consequences of what gets standardised in these shared resources.

cognitive science, artificial intelligence and neuroscience. While this conference is not explicitly focused on DNN research, its inaugural event centred around neural network research and the "mainstreaming of machine-learning techniques" (Naselaris et al., 2018, p.4). The DNN community within computational neuroscience is comparatively young, which explains its lack of development compared to animal model organism communities.

In summary, DNNs share some features typical to paradigm cases of model organisms. However, as I will show in the next section, DNNs do not fit the key features of wide representational scope and target. Therefore, I argue that it is more appropriate to interpret DNNs as Krogh organisms over model organisms.

4.3.2 Representational scope and target

Do DNNs share the commitment to representational scope and target typical of model organisms? Recall that representational scope concerns how extensively the results of research on an organism can be projected onto a wider group. While DNNs are meant to be generalisable, it is not clear what exactly their scope might be. In particular, it is not apparent that results from a DNN can be projected to a specific group such as a species given they do not belong to a well-defined group. To this extent, the representational scope is vague and limited by the fact that there is no shared evolutionary history to draw on with other organisms, a significant difference from animal model organisms, which have a well-defined contrast class (Scholte, 2018).

This feature supports my interpretation of DNNs as Krogh organisms characterized in terms of an open-ended representational scope. In DNN research, the representational scope is also treated as an empirical question where models are evaluated as to whether they may be generalised to different organisms' visual systems, such as mice and primates (Cadena et al., 2019). While

evolutionary history also has a role to play in phylogenetic inferences made from Krogh organisms, they can still be used to highlight differences with traditional models without shared evolutionary history.

With respect to representational target, Ankeny and Leonelli (2013) characterize model organisms as having broad targets because they aim for an integrative holistic understanding of an organism in terms of its development, physiology, evolution and other processes. While model organisms can still be used to address a narrower question about a particular phenomenon, research on them simultaneously serves the larger goal of building this integrative picture. In contrast, DNNs focus on narrow problem-focused targets, such as specific cognitive processes like object classification, unlike other modeling approaches such as neurorobotics, which arguably achieve a more integrative approach (cf. (Mitchinson et al., 2011; Prescott et al., 2009)). This further reinforces my interpretation of DNNs as Krogh organisms, which are not explicitly aimed at representing a target system but rather are problem-focused insofar as they aim to understand a particular biological problem.

In summary, DNNs lack two key features of model organisms. In particular, DNNs have a representational scope that is unestablished and their representational target is narrow. While DNNs may strive for model organism status in the future, I contend that currently DNNs are more similar to artificial Krogh organisms, exhibiting an open-ended representational scope, a narrow representational target, and are problem-focused rather than holistic.

4.4 What inferences can be drawn from artificial Krogh organisms?

Having argued that DNNs are artificial Krogh organisms, I now turn to the inferences that are drawn from these models and connect this to my discussion of Engineered Multiple Realization (EMR). As a reminder, EMR occurs when an agent engineers a system S that performs a function F in a way that is different from F's instantiation in a target system T. EMR is a perspectival account so what counts as "different" will depend on three components that constitute a perspective:

a) A notion of biological similarity or relevance B,

b) The methodology or technique for modelling the system M,

c) The assumed goals of the scientists G.

In other words, unlike traditional accounts of MR that view only one dimension as determining what constitutes MR, in EMR, scientists can avail themselves of MR along different dimensions depending on the three aforementioned components.

Why is EMR important? Because the process of engineering S aims to provide greater understanding of the target system T or the reproduced function F. In chapter 2, I argued that this can happen with recourse to two strategies. First, in order to gain greater understanding of a target system T, an iterative strategy is used where the guiding principle is to reduce MR in the dimension of choice. Second, in order to gain understanding of the function F, a comparative strategy is used where the guiding principle is to increase MR in the dimension of choice. Figure 1 provides an illustration of how these strategies relate to EMR. In chapter 3, I showed how these strategies are applied to particular cases, highlighting how both strategies allow for constraint-based reasoning by utilising specific methods, such as constraint mapping and constraint identification. In the

remainder of this section, I trace the connections between the framework offered in these past two chapters and how these different strategies rely on different uses of models.

4.4.1 Two modes of modelling

Bolker (2009) distinguishes between two modes of modelling in biology: surrogate and exemplary models. An exemplary model acts as a representative example of a species and thus relies on knowledge of phylogenetic context to judge its representativeness with respect to different features. A surrogate model acts as a proxy for a target where the model responds in the same way as the target would. Here the phylogenetic context is not necessary, all that is needed is for the model to accurately represent the process under study. Surrogate and exemplary models link to the two uses of EMR: 1) understanding the target system T, 2) understanding the function F. This, in turn, connects them to the iterative and comparative strategies respectively.

First, there are DNNs used as surrogate models, which rely on an iterative strategy to better match constraints onto formal constraints in models. This is exemplified by the performance-driven approach discussed in chapter 3 driven by the desire to perform artificial experiments *in silico* that are not viable on biological systems due to ethical concerns and the tools available (Kriegeskorte, 2015; Lindsay, 2020). One of the main goals of the performance-driven approach is to create a DNN that would serve as a proxy for a biological system to gain information about potential mechanisms linked to a particular behaviour (Tarr & Aminoff, 2016). In this case, computational neuroscientists have engineered multiple realization by producing a non-biological system that is able to achieve the levels of performance biological organisms have. For example, the production of DNNs for object recognition is an engineered multiple realization of object recognition. Stinson (2018) identifies one potential source of error for these kinds of models: the

output of the model is never exactly the same as the target phenomenon. Since surrogate models require the model to respond in the same way as the target system does, this would restrict the applicability of DNNs to biological systems. This issue is particularly pressing given that DNNs are not explicitly representing a particular target system in the first place. In order to act as a useful surrogate model, computational neuroscientists need to align their engineered system S with biological systems, that is, reduce the dimensions by which a DNN multiply realizes the kind object recognition. This contributes to how EMR can inform understanding of the target system T.

To achieve this, computational neuroscientists rely on the iterative strategy to make progress towards their ideal surrogate model. Through this iterative process, researchers find ways to minimize the differences between human and DNN output thus shifting towards making their EMR of a kind X "less" multiply realised. Recall the example of how DNNs were initially thought to have shape bias but then were found to actually detect visual texture in object classification instead (Geirhos et al., 2019). Rather than dismissing DNNs as a model of the human visual system, researchers then adjusted their datasets to force the network to rely less on texture thereby inducing a shape bias.[30] By using this iterative strategy, the hope is DNNs will eventually become more like holistic model organisms and contribute to understanding the mechanisms underlying object recognition. The iterative strategy used ultimately aims to make the realisers of a process as similar as possible helping support between-realizer inferences about a target system. By reducing the dimensions along which MR occurs, the model will be a more effective surrogate to use for uncovering structure-function relationships.

[30] Similarly, Peterson, Abbott, and Griffiths (2018) introduce a simple transformation to make neural network representations better approximate human ones.

Second, there are DNNs used as exemplary models, which rely on a comparative strategy that serves to identify potentially relevant constraints on a function F. This is exemplified, in part, by the anatomy-driven approach discussed in chapter 3. One of the main goals of this work is to compare lots of different models in order to gain information about the function you are interested in. In this case, computational neuroscientists engineer systems that perform a function in very different ways. This is where my account of EMR comes in. On Bolker's account, an exemplary animal model will be relevant as an exemplar of a species where you rely on phylogenetic inferences. In my account, we cannot think of a DNN as an exemplar of a species, instead we can interpret DNNs as one realizer of a function F. To gain information on constraints, a comparative approach is used, as discussed in Green et al. (2018), whereby systematic comparison of different species and the variations between different organisms helps the search for generalisations and design principles. In this case, the comparative method is applied across realisers such that the dimensions along which they vary becomes apparent. In this way, EMR can help us gain understanding of the function F.

To achieve this comparative approach in DNN research, methods that generate variability between models are required. While surrogate Krogh organisms use continuous iteration to minimize differences between model and target systems, exemplary Krogh organisms try to maximise the variation among instantiations of a particular behaviour in order to better guide the search for constraints that will generalise. While the iterative strategy has the elimination of MR as its regulative ideal, the comparative strategy instead tries to maximise the amount of MR in the dimension of interest. Thus, it generates a large number of possible alternatives to identify relevant constraints on systems that realize a kind or function. This strategy also is at work in the scientific research on DNNs. For example, recall how Musall et al. (2019) describe the ability to generate a

large number of possible instantiations of a particular behaviour as a benefit to computational modelling: “one can easily train thousands of networks while systematically varying the behavioral task, neural architecture and cost functions, thus enabling the study of large ensembles of potentially different solutions to a given behavioural task” (p.234). Having a large collection of diverse artificial Krogh organisms helps the comparative approach lead to the desired inferences by identifying constraints that determine performance. Constraint-based explanations in animals can show how a particular architecture guarantees a certain robustness (Green & Jones, 2016, p.362). For DNNs, this would amount to showing how certain features of a network are integral to performance on tasks. It is worth noting here that the comparative strategy cannot countenance maximization of MR on all dimensions, as this would make it difficult to know what to compare in the first place. For this reason, the standardisation we expect in model organisms will still be important for exemplary models as well.

In summary, the conception of DNNs as Krogh or model organisms is helpful for understanding the different epistemic uses of these models as they function in EMR. In EMR, DNNs are generated to support two strategies. As a surrogate model, which is more similar to a model organism, DNNs would support the iterative strategy. As an exemplary model, which is more similar to a Krogh organism, DNNs support the comparative strategy. If I am right that DNNs are currently more like Krogh organisms, then they are currently more justified for use as exemplary models than as surrogate models.

4.5 Objections and responses

Having spelt out my account of DNNs as artificial Krogh organisms, I finally turn to responding to other proposals made for how DNNs could be models and address why they are insufficient or compatible with my account. I also consider some objections to my proposal.

4.5.1 Other philosophical accounts of models

First of all, DNNs could be identified as belonging to a traditional kind of theoretical model in computational neuroscience, such as the Hodgkin-Huxley model (Cichy & Kaiser, 2019). In theoretical models, scientists construct a simplified mathematical or mechanistic version of a target system that can be analysed and potentially lead to conclusions about the target (Levy & Currie, 2015). For example, the Hodgkin-Huxley model is a mathematical model consisting of a set of differential equations that accurately predicts several features of action potentials by providing an abstracted and simplified description of the process (Levy, 2014). With their highly abstract mathematical descriptions, it is tempting to analyse DNNs as theoretical or mathematical models.

Here one might ask: what difference does it make how a model is categorized and what is so distinctive about calling something a model organism? Philosophers are divided on whether to make a strict demarcation between model organisms and other kinds of model. On one side, there are some who blur the distinction between model organisms and other categories of model. For example, Weisberg (2012) subsumes model organisms into the category of concrete models. Ankeny and Leonelli (2011, 2020) situate model organisms within Morrison and Morgan's (1999) ‘model as mediators’ framework. On these views, though model organisms may have special unique features compared to other models, they still bear important similarities to theoretical

models collapsing the distinction. If one agrees with this perspective, my argument for understanding DNNs as model organisms and Krogh organisms need not invalidate interpretations that they are also theoretical models. In this case, my argument would suggest that they are a special kind of theoretical or concrete model.

On the other side, there are some who defend a strict demarcation between model organisms compared to other categories of model. For example, Levy and Currie (2015) propose a distinction where model organisms and theoretical models differ based on the strategies they use to justify inferences from model to target. They suggest that theoretical models are surrogates, relying on analogies between the model and target. In contrast, inferences from model organisms are to be thought of as empirical extrapolations based on shared ancestry and empirical information. This distinction does not strike me as successful in demarcating between kinds of models. Arguably model organisms can function as surrogates too – a point that Levy and Currie (2015) themselves appear to concede and that is forcefully made by Parkkinen (2017) through case studies of the use of model organisms in medicine. Parkkinen (2017) also argues that model organisms are epistemically distinct from theoretical models. However, instead of basing the distinction on how model-to target-inferences are justified, Parkkinen identifies a different role that theoretical models play. This role distinguishes them from model organisms based on the different epistemic purposes of the models. The epistemic purpose of theoretical models is to explicitly formalize assumptions about the target system such that no further analogy is needed to determine its relevance to the target system, as long as these assumptions are empirically adequate. In contrast, animal models such as model organisms are used as a surrogate that can provide evidence for hypotheses about a target system. Here, one must endeavour to establish similarities between the model and target system in order to extrapolate to the target system.

Even if one accepts Parkkinen's (2017) account of the difference between theoretical models and model organisms, I contend that DNNs are still more similar to model organisms than theoretical models. This is because DNNs are typically not explicitly formalizing assumptions about a target system. While DNNs are often described as inspired by biological neural systems, the analogy is loose and DNNs are not explicitly aiming to represent the brain as a target system.[31] The design of DNNs typically consists of three components: the objective functions, the learning rules, and the architecture (Richards et al., 2019). It is controversial to claim that the objective functions and learning rules used in DNNs are the same as in the brain (Lindsay, 2020; Marblestone et al., 2016). For example, back-propagation, the learning rule used in most deep learning, has long been notorious for its biological implausibility even from when it was first used in early connectionist networks (Crick, 1989). This leaves us with architecture, which is commonly appealed to as biologically inspired, however, even then, the choice of architecture is not separate from task considerations (Hasson et al., 2020, p.429). Given this, one can conclude that the model to target relation in DNNs is an ambiguous one and that the design of DNNs is not explicitly aiming to represent a biological system. Instead, DNNs bear hallmarks of the process described by Parkkinen, whereby the similarities between the model and target system must be established after the fact. We can see attempts to do this by DNN researchers who work to demonstrate how back-propagation can be biologically implemented or plausible (Lillicrap, Cownden, et al., 2016; Lillicrap, Santoro, et al., 2020). These practices are similar to what Ankeny and Leonelli (2020) describe as 'plausibility building' where researchers work to demonstrate a model is plausible. If DNNs were theoretical models based on Parkkinen's account, then this process would be

[31] Also see Weisberg's (2012) discussion of targetless models.

unnecessary. It is because DNNs must have their utility and representational power demonstrated, just like animal models, that computational neuroscientists are concerned with showing they are biologically plausible.[32]

I argue that the notion of plausibility that Ankeny and Leonelli (2020) introduce gives more reason to view DNNs as similar to model organisms and Krogh organisms. Ankeny and Leonelli introduce plausibility as a measure of when a model organism counts as a good model. They recognise biological plausibility as a sociological phenomenon: "what makes an organism-based model plausible as a representation depends on the degree to which communities of researchers deem the use of the organism as a model for a given phenomenon to be epistemically fruitful and justifiable within the broader research environment" (2020, p.52). First, this is a dynamic concept of plausibility, where a model can gain or lose plausibility over time. This fits in well with the practices that we observe in DNN research where one interpretation of their performance can make them seem more plausible but later interpretations may reduce their plausibility. Second, this concept of plausibility will be broader than a mechanistic or evidence-based approach to determining plausibility since this notion also gives us insight into why and when model organisms are viewed as more or less useful to scientific practice.

A key point in Ankeny and Leonelli's account of plausibility is that it introduces the process of 'plausibility building'. Since plausibility is a dynamic process, researchers have a motivation to try and build up the plausibility of their model over time. For example, Ankeny and

[32] A consequence of this is that we might reconsider the importance of interpretability for these models by wondering whether model organisms must be interpretable in this way too. For animal models, we require experiments to be performed to gain understanding. Similarly, the same neuroscientific methods such as lesioning and anatomical tracing can be used in silico on DNNs to make them more interpretable (Lindsay, 2020).

Leonelli suggest this happens in the use of mouse models for alcoholism research (Ankeny & Leonelli, 2020; Leonelli et al., 2014). Researchers standardised certain key features that would represent alcoholism in order to be able to extrapolate mouse model results to human alcoholism. These methods and features made the models more plausible despite initial reservations about the applicability of mouse models to this phenomenon. A similar process exists in research with DNNs. Researchers need to work to demonstrate that DNNs are good representations of biological systems and they do this by invoking particular notions of biological plausibility such as ecological validity, representational similarity or performance similarities. This can be supplemented by my perspectival account of EMR. Since what counts as biologically similar is dependent on a researchers' perspective, researchers may often call on other notions of biological similarity, informed by other perspectives to further reinforce the plausibility of their model. Indeed, this is one way we might interpret the scientific interest in demonstrating how back-propagation could be implemented in the brain. Though this does not concern those who use DNNs as models of the brain to investigate performance, these researchers "borrow" the notion of biological similarity from other researchers who care about implementation issues to solidify the biological plausibility of their model. However, since this is a sociological process, this will often be a negotiation that requires different perspectives to compromise on what they take to be sufficient to fulfil their notions of biological similarity.

4.5.2 Other scientific accounts of DNNs

I have framed my argument around the frequent discussions by scientists describing DNNs as artificial model organisms or experimental organisms. However, there are other ways scientists

talk about DNN models as well. Here I consider these other proposals and compare them with my account.

Kell and McDermott (2019) propose DNNs should be analysed as similar to ideal observer models, or models typically derived to determine optimal performance in a task based on certain specifications (Geisler, 2001). As such, a DNN can then be compared to the biological system to act as a benchmark for how a task should be performed. Kell and McDermott (2019) accept that there are dissimilarities: DNNs are not provably optimal unlike traditional ideal observer models. However, there are further issues for the ideal observer model when applied to DNNs. First, DNNs are not used as a benchmark for human neural systems, rather vice versa: human behaviour and neural predictivity act as benchmarks for the success of DNNs instead of DNNs showing how biological systems ought to act. Furthermore, the analogy between a DNN and an organism depends on the idea that since the optimization algorithms in the brain have been subject to evolutionary pressures, the brain must have found a (locally) optimal solution. Hence the utility of DNNs can be justified if they are also optimal (Marblestone et al., 2016). If DNNs are not providing optimal solutions, then their utility on the basis of this argument would not be justified.

A similar issue applies to a proposal advanced by Hasson et al. (2020), which draws on an analogy to evolutionary theory. Namely that DNNs are 'direct-fit models' which utilize brute-force optimization over distilled design principles. Now, whilst evolution may not in fact commit us to saying that organisms are optimally adapted (Parker & Smith, 1990), the burden of proof is on the direct-fit accounts to spell out the purported analogy between the 'evolution' of biological and artificial systems in these accounts. While I dismiss treating DNNs as ideal observer models, I agree with Kell and McDermott (2019) on the importance of DNN models for revealing task constraints, which is compatible with my account of DNNs as Krogh organisms. One issue is that

Hasson et al.'s proposal would reject the view that DNNs give us insight into design principles. However, as I discussed earlier, this would still leave DNNs open for use as surrogate models or exemplary models that help us learn about a target system T.

Kay (2018) makes an important distinction between functional and mechanistic models,[33] where functional models capture the input-output descriptions of a system and mechanistic models try to use parts that correspond to the actual parts of the target system. Lindsay (2020) interprets DNNs as mechanistic models, though I argue that DNNs are probably closer to functional or phenomenological models striving to be mechanistic ones. In the next chapter, I elaborate on why I think we should reject mechanistic interpretations of DNNs by focusing on arguments by Cao and Yamins (ms).

Cichy and Kaiser (2019) emphasise the importance of treating DNNs as exploratory models, which can generate new hypotheses about how the brain works; demonstrate the feasibility of a particular approach to problem-solving; and assess our characterization of a phenomenon. My proposal is compatible with the exploratory power of DNNs and indeed I think constraint-based reasoning is a helpful way to make sense of how exploratory models provide information. Exploration can help us understand the role of constraints on a phenomenon, which achieves all three of Cichy and Kaiser's goals for exploratory models. First, there is the generation of new hypotheses about how the brain works. The engineering of particular solutions to a problem the brain also faces can provide inspiration for how the brain might achieve it as well. Second, constraint-based reasoning can permit us to demonstrate how a particular approach solves a problem. As discussed earlier, the use of the comparative approach serves to identify constraints

[33] This echoes an earlier distinction by Luce (1995) between phenomenological and process models, where the former treat a phenomenon as a black box in order to describe its patterns of behaviour without the internal structure, and process models, which attempt to open the black box by modelling the internal mechanism.

connected to particular phenomena and in turn influences our characterization of phenomena or forces us to adapt them. Krogh organisms provide a good framework for making sense of what we gain from exploratory models.

4.5.3 Further objections

4.5.3.1 Objection 1: DNNs are not organisms!

A natural response to my argument in this chapter is to take issue with the comparison between DNNs and animal models by highlighting a key difference between the two and argue that DNNs are not organisms at all. Ankeny and Leonelli (2020, p.19) argue that model organisms cannot be built from scratch in a laboratory because we understand only part of how they work so we could never hope to reproduce them. They consider the model organism's hybrid status, as both natural and artifact, to be key to their use. Computational models like DNNs clearly do not share this hybrid status. While this is related to their account of model organisms, this objection could still hold for my interpretation of DNNs as Krogh organisms as well. Here I offer a few responses to this line of argument based on their account.

First, I contend that model organisms are partial representations and Krogh organisms even more so, since they focus on a specific problem rather than a holistic account of an organism. In this case, we would not need complete reproduction in a laboratory to fulfill the model's epistemic role. Ankeny and Leonelli (2020) discuss how the representation we get from model organisms is not a static mirroring relation. Rather, it is a dynamic process, which is necessarily partial since a model organism's representational power comes from research practices, which establish and justify its ability to generalize. It may be the case that model organisms *qua* natural objects offer

more validity 'for free' than DNNs. However, as I will expand upon later, model organisms also require a significant amount of engineering in order to allow us to make inferences to target systems. What this suggests is that the process of establishing relevance and 'plausibility building' will be particularly important for DNNs.

Second, I contend that an important reason for the 'natural' component of model organisms is due to their use as genetic tools. However, I view this as a contingent rather than essential feature of model organisms. This puts me at odds with Ankeny and Leonelli who emphasise genetic standardisation as an important feature for justifying the use of model organisms. However, Ankeny and Leonelli point out that this is partly due to historical reasons: "a genetically based approach to understanding cross-species comparison and in turn standardisation was not strictly necessary for the conceptualization of the category of model organisms and their use […] for reasons that were at least partly contingent, the classical tradition of genetic analysis ended up playing an important role" (2020, p.15). If the conceptualization of model organisms does not require genetic analysis, then it is possible to conceive of another way to think of model organisms divorced from their natural genetic importance. This brings me to my last point.

Third, Ankeny and Leonelli's (2020) account focuses mainly on biological cases to support their view of model organisms. Perhaps the conception of model organisms in cognitive science, especially that of a computational flavour, is different and grounded in other practices. One reason to think this is that, while biology relies on a stricter natural/artificial distinction, cognitive scientists frequently seem comfortable collapsing or blurring this distinction. As I discussed in chapter 2, computational neuroscience has its own set of assumptions that license this, namely the belief that both computers and brains can be analysed solely in terms of information processing, thus abstracting away from lower-level details.

If one still remains unconvinced on this issue, then I note that Ankeny and Leonelli also suggest that model organisms need not be the only scientific models that will have the key features of wide representational scope and target (2020, p.12). To this end, if you want to reject thinking of DNNs as model organisms or Krogh organisms on the basis that they are not organisms, then you can take my argument as suggesting that DNNs are a kind of model that display the same key characteristics of model and Krogh organisms with regards to representational scope and target.

4.5.3.2 Objection 2: Model/Krogh organisms are not engineered!

Another concern is that a key difference between model or Krogh organisms and their artificial counterparts is that the animal models are not engineered. However, I contend that there are many ways in which animal models, particularly those used as model or Krogh organisms, are, in fact, engineered as well.

First of all, there is a kind of pragmatic engineering, namely the way in which selection of animals will depend on a variety of pragmatic factors to make experimenting on them easier. Dietrich et al. (2019) provide a set of twenty criteria that guide organism choice, which shows the selection of particular organisms is not a coincidence but partly dependent on researcher goals. In particular, animal model organisms will not necessarily be picked based on what is most representative of human behavioural or neural patterns. Rather, other factors such as what makes it easiest to investigate particular phenomenon will play a determining role.

Second, genetic standardisation is a hallmark of many model organisms, resulting in a standard strain used for research that establishes little variability across experiments.[34] Many features of model organisms are the product of human intervention, as they are developed to possess features valued by researchers. This too involves a form of engineering. Leonelli (2007) traces this process of standardisation for one of the most popular plant model organisms, *Arabidopsis thaliana*. Ankeny and Leonelli are even more explicit: "though of course model organisms have their origins in the wild, they are in fact constructed through a diverse range of practices" (2020, p.11). A key point in these analyses is that the preparation required to make model organisms useful for research makes them into technical artifacts.

Another kind of standardisation arises through behavioural engineering, or the modification of behaviour to accommodate experimental investigation. I discussed how this applies to DNN research but it also features in animal model research too. The artificiality of laboratory environments has long been recognised as imposing limits on ecological validity. Canguilhem notes,

> "we must not forget that the laboratory itself constitutes a new environment in which life certainly establishes norms whose extrapolation does not work without risk when removed from the conditions to which these norms relate[…] it is not possible that the ways of life

[34] Ankeny and Leonelli: "more specifically, model organisms have particular experimental characteristics that are closely related to their power as genetic tools: they typically have small physical and genomic sizes, short generation times, short life cycles, high fertility rates, and often high mutation rates or high susceptibility to simple techniques for genetic modification. Furthermore, they have been developed using complex processes of standardisation that allow the establishment of a standard strain which then serves as the basis of future research. The standard strain, often paradoxically referred to as "wild type," is a token organism developed through various laboratory techniques (ranging from cross-breeding to genetic manipulation) so that it possesses features valued by researchers and can be reproduced with the least possible variability across generations, for example through cloning." (2012, p.13)

in the laboratory fail to retain any specificity in their relationship to the place and moment of the experiment" (1989, p.85).

One way in which laboratory norms are imposed is in the modification of the environmental context that animals find themselves in. Indeed Ankeny and Leonelli contend that the standardisation of a model organism's environment that is achieved in the laboratory is the "biggest source of uniformity" (2020, p.24). Concerns about the artificiality of laboratory of environments have been raised by cognitive ethologists who challenge that laboratory experiments can directly provide information about natural behaviour in real-world scenarios. For example, Kingstone et al. (2008) point to two underlying assumptions of laboratory research: the invariance assumption, which assumes that cognition is invariant and regular across different situations and the control assumption, which assumes that one can strip context away from laboratory experiments without compromising what one is measuring. They argue that the principle of invariance cannot and should not be assumed since "based on laboratory findings alone, it is not possible to know whether mechanisms that appear invariant in the laboratory environment will survive outside the lab" (ibid., p.319). They call for the need to first study how people behave in natural environments before moving investigation into the lab.

The proposals for caution regarding inferences made from laboratory experiments is reinforced by research showing that there are behavioural differences between animals in the laboratory and in the wild. Calisi and Bentley (2009) discuss ways in which laboratory animals may not be the "same" animal as those in the wild. They focus on how the laboratory environment can especially affect experiments on endocrinology and behaviour. For example, changes in diet and social dynamics of the laboratory animals affects brain morphology and the behaviour of animals. As Kingstone et al. (2008) argue, scientists have reason to give up the invariance

assumption that animals are acting exactly as they would in the wild. Overall, this work raises important questions about what can be extrapolated from laboratory experiments alone and whether this may pose challenges for the representational power of model organisms. However, regardless of the answer to those questions, this demonstrates that there are significant ways in which there is behavioural modification in laboratory environments so behavioural engineering occurs for animal models too. In summary, there are important ways that animal models used in research as model organisms or Krogh organisms are engineered.

One consequence of establishing animal models as engineered is that it may challenge the application of standard accounts of multiple realization to these cases. Instead, this gives further reason to apply EMR to conceptualise this research as scientists engineering animal model systems to realize a function. The engineering of model organisms fits this narrative. Scientists treat animal models in ways that they see as better placed for generating scientific knowledge even if, in doing so, the animal models do not really represent how things occur in the wild.

4.6 Conclusion

In this paper, I sought to answer two questions regarding deep neural networks. First, what kind of model is a DNN? I concluded that DNNs are artificial Krogh organisms, whereby their open-ended representational scope is treated as an empirical question in scientific work. Second, what kind of inferences can be drawn from these DNN models? To answer this, I connected two epistemic strategies for exemplary and surrogate Krogh organisms and showed how they support constraint-based inferences as well as reflecting on their limitations. While there are still important distinctions to be made between animal model organisms and artificial counterparts, I point to

significant parallels that suggest how neuroscientists treat animal model organisms and Krogh organisms can provide lessons for how to understand what DNNs do for computational neuroscientists. Furthermore, I related Krogh organisms to the EMR framework I provided in the previous chapter to show how the use of model organism is better understood in the framework of EMR rather than in traditional multiple realization frameworks. This could undermine some of the ways animal model research is used to support ascriptions of traditional multiple realization. To conclude, my account provides a strong motivation for why computational neuroscientists engineer artificial organisms and offers further insight into the role that DNNs play within neuroscience.

5.0 A criticism of (some) mechanistic interpretations of DNNs in computational neuroscience

> *Current neurobiology is like somebody coming in with shelves and shelves of transistor manuals and saying, "See, this is how, this is what the world reduces to." And you say, "But tell me how to put them together." […] whatever the biologists are discovering tells you nothing about the nature of the system. I mean, the components don't tell you anything about what the system does, only how it works. So, from my point of view, I'm a strong supporter even if I'm not a sympathizer or participant in artificial intelligence. Forget about biological preoccupation with receptors and transmitters and magic molecules. You take that for granted as mechanism, but the process of a system is a different thing from its mechanism.*
> ~Lettvin (2000), p.15

In this chapter, I survey the current debate on mechanistic explanation and multiple realizability in computational neuroscience. In particular, I argue against some recent proposals interpreting DNNs as providing mechanistic explanations. In section 1, I provide an overview of the literature on mechanistic explanation in computational neuroscience and the role that multiple realization has played in these discussions. In section 2, I identify how the denial of multiple realization is a key premise in Cao and Yamins' argument for a mechanistic interpretation of DNN models. In section 3, I present two forms of MR that Cao and Yamins' may wish to deny for DNNs and show that both cases do, in fact, appear to be multiply realized, according to my account of EMR. In section 4, I argue the reason that the denial of MR fails in this case is due to an underdetermination argument that can be made for current DNN research, which means that we can't identify what features are most relevant for the success of DNNs. I end by discussing why this underdetermination argument still allows us to engage in constraint-based reasoning and explanation showing that DNNs can provide non-mechanistic explanations. I conclude that my

argument gives us reason to re-evaluate the emerging consensus in the literature that DNNs are mechanistic models.

5.1 Mechanist interpretations of DNNs in neuroscience

Mechanistic explanation has established itself as a dominant framework for explanation in neuroscience (Craver, 2007; Kaplan & Craver, 2011; Machamer, Darden, & Craver, 2000; Milkowski, 2013). While there is discussion of whether computation itself can be understood in mechanistic terms (Piccinini, 2015), here I am mainly concerned with how models are used in computational neuroscience and whether these models provide mechanistic explanations. This is a topic of ongoing debate.

The standard way that mechanists make sense of how models can provide mechanistic explanations is by proposing a model-to-mechanism-mapping, otherwise known as 3M (Kaplan, 2011; Kaplan & Craver, 2011).[35] 3M states that for a model to provide a successful explanation:

a) "The variables in the model correspond to components, activities, properties, and organizational features of the target mechanism that produces, maintains, or underlies the phenomenon, and

b) The (perhaps mathematical) dependencies posited among these variables in the model correspond to the (perhaps quantifiable) causal relations among the components of the target mechanism" (Kaplan, 2011, p.347).

[35] Though see also Hochstein (2016) for a useful discussion on two different understandings of mechanistic explanation in the literature and how they apply to modelling.

Some have pointed to specific examples of computational models in neuroscience to argue that they resist a mechanistic analysis. Chirimuuta (2014, 2018) discusses various examples of research in computational neuroscience that are resistant to mechanistic interpretations. Weiskopf (2016) uses the case of integrative cognitive models to argue against a mechanistic analysis, endorsing a model pluralism and dispensing with the assumption that there is a unique causal structure. Mechanists have responded defending mechanistic explanation against these challenges (Bantegnie, 2017; Kaplan, 2017).

In one of these responses, Kaplan (2017) argues that the legitimacy of mechanistic explanation in computational neuroscience is not threatened by multiple realizability. Kaplan's target here is the purported multiple realization of Canonical Neural Computations (CNCs) discussed in Chirimuuta (2014). CNCs are neural computations that are repeated and combined in different ways across the brain, with linear filtering and normalization serving as illustrative examples (Carandini, 2012; Carandini & Heeger, 2012). Since these CNCs are not closely linked to biophysics, they are "less likely to map one-to-one onto a biophysical circuit" (Carandini, 2012, p.508). Thus CNCs could be implemented by different biophysical mechanisms making them multiply realized (Chirimuuta, 2014).[36] Chirimuuta argues that CNCs do not sit well with the central commitments of the mechanistic approach. First, because CNCs are models that do not map onto components of a mechanism, thereby failing to satisfy 3M. Second, because CNCs show that the assumption that "more details are better" (MDB) need not apply for computational explanations. In particular, a mechanistic interpretation will fail to capture the explanatory power

[36] It is worth noting that Chirimuuta (2014) makes clear she endorses a weak version of MR, where "computational properties of neural systems… appear to be partially independent of their mechanistic realizers" (148) for CNCs. This remark suggests an implicit perspectivism of MR, as was highlighted in §2.3.

of CNCs: “any digging down to more mechanistic detail would simply lead us to miss the presence of CNC’s entirely, because of their many different realizations” (Chirimuuta, 2014, p.140). Because of this, Chirimuuta argues that CNCs fail to be accommodated by the mechanistic framework and should be understood as minimal model explanations.

In his response to Chirimuuta, Kaplan (2017) argues that MR considerations are not relevant to assessing the adequacy of a mechanistic explanation. He interprets a key premise of Chirimuuta’s argument to be:

> *(M) “If a given phenomenon is multiply realizable, then a model of that phenomenon will not satisfy 3M” (p.173).*

Kaplan argues that (M) is false since MR considerations are not relevant to assessing the adequacy of a mechanistic explanation – a multiply realized phenomenon can still satisfy 3M and therefore warrant mechanistic models. To understand this better, let us consider the details of Kaplan’s argument. In abstract, Kaplan distinguishes between narrow and wide scope mechanistic explanations. Narrow mechanistic explanations are local and thus may just be explanatory of a particular realization of a phenomenon. In contrast, wide-scope explanations are more general and unify all known systems realizing that phenomenon. He provides examples of two different mechanistic models of sound localisation in birds and mammals. Then, he argues that each of these mechanistic models individually satisfy 3M and thus are adequate mechanistic explanations of the multiply realized phenomenon of sound localisation. They fail to be wide-scope explanations since neither one of these mechanistic models can subsume or explain the phenomenon of sound localisation in the other species. However, they both succeed in being local narrow-scope mechanistic explanations for birds and mammals respectively. Thus, Kaplan claims we can dismiss (M) since, even though sound localisation is multiply realized, we can still develop models of

sound localisation that satisfy the model-to-mechanism condition and provide narrow mechanistic explanations.[37]

While Kaplan's argument does not fully address Chirimuuta's argument based on CNCs (since she does not take MR to be the only reason that CNCs fail to be mechanistic), his paper still serves to illustrate one perspective a mechanist could take on whether multiple realization affects mechanistic interpretations of computational models. However, I contend that Kaplan's argument appears to leave room to accept multiple realization and non-mechanistic models. Based on Kaplan's terms, even if we grant that there could be local narrow-scope mechanistic models of a multiply realized phenomenon, this still does not account for the explanatory force of the CNC that is multiply realized. In denying (M), Kaplan has shown that if a given phenomenon is multiply realizable, there can still be *some* mechanistic models of the phenomenon that satisfy 3M. But he seems to suggest that anything that satisfies 3M here will be a narrow-scope mechanistic model. This does not eliminate the possibility that there are wide-scope models, albeit models that don't satisfy 3M, that could feature in non-mechanistic explanations. Perhaps anticipating this line of argument, Kaplan argues that we should not *prioritise* wide-scope explanations over the narrow-scope mechanistic ones. However, we need not view the wide-scope explanation as a "better" explanation in order to acknowledge that it is providing a different explanatory role to the narrow-scope mechanistic explanations. This is reflected in Dewhurst & Lee's (2021) analysis of the disagreement between Chirimuuta and Kaplan. They claim the disagreement can be dissolved by interpreting the interlocutors as adopting different stances: Chirimuuta is adopting a design stance,

[37] Note that another potential route for the mechanist is to argue that computational explanations are mechanism sketches (Piccinini 2015). However, see Polger and Shapiro (2016, p.160) for some commentary on the tension between the view that computational explanations are mechanism sketches and the view that computation is realizer-independent.

asking what a computation is *for*, while Kaplan adopts a mechanistic stance, asking *how* computations realize their function. In summary, I claim it is still legitimate to ask after the explanatory purpose of the wide-scope models of a multiply realized phenomenon, which is what I believe some philosophers think of Deep Neural Networks.

The use of DNNs in neuroscience has not been immune to mechanistic readings and, though philosophical work on this is in early stages, there seems to be a fast-emerging consensus on a mechanistic interpretation (Buckner, 2018; Cao & Yamins, ms; Stinson, 2018, 2020, Ritchie, ms). In this paper, I argue against this trend by specifically charting how multiple realization is being invoked in this debate. In contrast to Kaplan, I claim that the denial of multiple realization plays an important role in some recent arguments for interpreting DNNs as providing mechanistic explanations, precisely because DNNs aim for wide-scope explanations that attempt to unify different systems that realize a phenomenon. I focus on Cao and Yamins as a central example to illustrate my argument. In Cao & Yamins (ms), they have two aims:

a) Propose a new model-to-mechanism-mapping 3M++ to show how DNNs provide mechanistic explanations contra claims that DNNs lack biological realism,
b) Demonstrate how constraint-based reasoning makes DNNs intelligible contra claims that DNNs lack interpretability.

While these aims can be understood separately, Cao and Yamins also combine them by presenting a no-miracles argument for the explanatory capacities of DNNs to match structure to function. In response to concerns about the lack of interpretability of DNNs, Cao and Yamins argue that we can understand these models through constraints that place restrictions on the types of possible solutions to difficult problems. They propose two forms of constraints: behavioural constraints and architectural constraints. Then, Cao and Yamins use constraints as a way to justify

a mechanistic mapping relation with analogy to the no-miracles argument, "if the constraints[…] were completely wrong, it would be a kind of *miracle* for the model satisfying those constraints to nonetheless get the predictions for intermediate-level neural activity right" (p.18). Cao and Yamins propose, since the model's success is not a miracle, then the successful prediction is due to similarity between the DNN models and brain regions responsible for the same functional capacities. In summary, the predictive success of DNNs is taken to be evidence for the mapping required for mechanistic explanation.

Unlike the examples that Kaplan concerns himself with, Cao & Yamins are engaged in the project of defending DNNs as both wide-scope explanations and mechanistic explanations. In other words, the DNN is not just a model of human object recognition but a generic model of object recognition applying to all or several object recognition systems. This aligns with both Buckner (2018) and Stinson (2018, 2020) who interpret neural network models in neuroscience as generic or idealised mechanisms. I contend that, once we are talking about wide-scope explanations, Kaplan's argument no longer holds. In fact, as I will show in the next section, Cao & Yamins' argument relies on claims about multiple realization in order to succeed.

5.2 Multiple realization and the no-miracles argument

As I outlined in §5.1, Cao and Yamins present a no-miracles argument to defend their mechanistic mapping relation between DNNs and the brain. In this section, I will demonstrate how this argument relies on a denial of MRT – the multiple realization thesis. In other words, the no-miracles argument that supports a mechanistic reading of DNNs relies on the realized function of object classification not being multiply realizable or, at least, having very few realizations.

Intuitively, this link between no-miracles arguments and MRT is not surprising since no-miracles arguments seem to only work if you think there are very few ways for the outcome taken to be a "miracle" to occur. For example, consider a case where you find yourself at a new restaurant you've never been to before and you order a burger. As you start eating, your friend asks you what you think the burger is made of. Since you think the burger has the same appearance, taste and texture as every burger you have had made of meat in the past, you exclaim it would be *a miracle* if this burger wasn't made of meat too, and you conclude that it is made of meat. As it happens, unknown to you, the restaurant you are at is a vegan restaurant and the burger is not made of meat but one of many vegan meat substitutes. Say you knew that the restaurant was vegan and were well-versed in the various meat substitutes that still had the same appearance, taste and texture of meat, the appeal of the no-miracles argument would be substantially diminished. In that scenario, you wouldn't consider the burger having the same appearance, taste and texture of meat as reason to conclude that the burger must be made of meat because you would know the same taste and texture can be multiply realized by non-meat alternatives. Once you are aware of this multiple realization, then the no-miracles argument doesn't work. It is only if you are in a situation where you believe the texture and taste is not multiply realized that the no-miracles argument would be convincing.

A similar assumption is underlying Cao and Yamins' no-miracles argument. They suggest that with respect to tasks like object recognition, "we are closer to the first extreme… where there are few solutions, or perhaps only one, so that systems that succeed at the task are forced to have certain features" (p.18). Cao and Yamins justify this assumption by appealing to the tasks the DNNs must solve as "difficult predictions". I reconstruct their argument as the following:

P1) The human brain can solve tasks such as object recognition.

P2) Object recognition is a "difficult" prediction task, which can only be solved in a very small number of ways, i.e.) object recognition is not multiply realized.

P3) Some DNNs can solve object recognition tasks.

P4) Given P2, it is reasonable to assume that humans and DNNs are solving the task in a mechanistically similar way.

C) Thus there will be a model-to-mechanism mapping between brains and DNNs.

Premise 2 – the premise that denies multiple realization - is crucial for their argument. They claim that the task of object recognition is a difficult prediction task. So in order to solve a difficult task, it is expected that fulfilling a functional goal would end up determining other parts of the system, in this case, securing the model-to-mechanism mapping. If we weren't in the extreme case with few solutions and instead there were several solutions to the problem (i.e., if there was multiple realization), then the inference to a mapping between model and mechanism would no longer be secure. Just as I suggested with the vegan burger example, you need the denial of multiple realization to get this no-miracles argument to work.

Though Cao and Yamins never phrase their argument in terms of multiple realization, their premise 2 is analogous to Shapiro's (2004) mutual constraint thesis (MCT), which he offers as a competing hypothesis to multiple realization. MCT is the thesis that "there are few distinct kinds of brains that can actually produce human-like psychological capacities" (p.106). To support this, Shapiro (2004) argues that human minds are subject to several constraints that limit the amount of phenotypic variation across realizers. The consequence of this is convergent evolution where there will be few structures that realize a particular function. He emphasises that "the more complex a system [is], the more constraints it will face" (p.86) – where the human brain is taken to be one such complex system. Cao and Yamins' framing of premise 2 is very similar: "having more

constraints (again, all else equal) means fewer systems that satisfy those constraints[…] If it turns out that our object categorization task is difficult, then it is likely that any system that can successfully perform will have just the right set of features, arranged in just the right ways." (p.22) Cao & Yamins even explicitly describe this as "something similar [to] convergent evolution in nature" (p.22) – again convergent evolution is taken to be evidence against MRT. Since Cao & Yamins' P2 commits them to a form of MCT, then it will also commit them to a denial of MRT based on Shapiro's characterisation.

Here one might respond that Cao & Yamins are not talking about MRT in the same way that Shapiro is since the latter is discussing implementation, while Cao & Yamins are talking about differences about computational strategies. However, as I argued in chapter 2, the traditional understanding of multiple realization purely in terms of implementation is inadequate to account for the scientific practice of engineering realizers of a function. Instead, I offered an explicitly perspectival account of Engineered Multiple Realization (EMR), where multiple realization depends on the notion of biological similarity being used, the methodology for modelling the system and the assumed goals of successful EMR. This perspectivism means that there are different forms of MR that centre different kinds of biological similarity. On this account, we can interpret Cao & Yamins as referring to particular forms of MR that do not focus on implementation as the central kind of biological similarity. Perhaps one could reject my account of EMR as applicable here. However, I think that aspects of EMR are already implicit in Cao and Yamins' paper. Early on, they note, "to suppose any model system (even a biological one) could be similar to a target brain is to have already accepted that some differences may be safely ignored" (p.4). For the purposes of their discussion, they are prioritising certain constraints and forms of biological similarity but they admit there are others one could prioritise. For example, the fact that DNNs and

brains are made of different materials means there are metabolic constraints restricting neural architecture, or that brains are subject to evolution means that there are historical developmental constraints DNNs won't be subject to. Thus, there is implicit recognition that both scientists and philosophers pick what kinds of multiple realization are relevant for the investigation or argument at hand. In order to evaluate Cao & Yamins' denial of MRT in this case, we need to specify what forms of MR are relevant to support their argument.

5.3 Why we can't deny MR for DNNs

In this section, I establish two forms of MR as the targets of Cao & Yamins' argument and consider each in turn to show there isn't clear evidence in favour of denying either form of MR. As I discussed in chapter 3, the use of DNNs falls under what I call the performance-driven approach. On this approach, the initial biologically relevant feature is to assess the performance and behavioural features of a model. From there, researchers then look to connect these models to other biologically similar features, such as by observing whether there is representational similarity between the brain and DNNs. For Cao & Yamins' argument, they are relying on the lack of MR with respect to behaviour and representations (i.e. claiming that DNNs and brains perform functions in similar ways and use similar representations) in order to bolster their conclusion of mechanistic similarity between brains and DNNs. Here I provide evidence against the similarity of DNNs and brains in terms of behaviour and representations.

First, I consider behaviour – are DNNs and brains behaviourally similar? In Cao & Yamins' paper, they discuss one biological constraint that DNNs adhere to – a behavioural constraint based on object recognition performance. If humans and DNNs perform object recognition in similar

ways, then it supports Cao & Yamins' denial of MR and thus P2 in their no-miracles argument. However, this behavioural constraint could be interpreted in two different ways: i) an accuracy threshold in attaining similar levels of correct categorisation as humans or ii) as DNNs exhibiting similar categorisation strategies as humans. There is evidence of i) since one could argue DNNs do meet accuracy thresholds in object categorisation similar to those of humans. However, I take it that Cao & Yamins would need ii) in order to show there is no MR at the behavioural level. But we cannot take evidence of i) as evidence for ii). As Geirhos, Meding, & Wichmann (2020) show, i) and ii) come apart. In their paper, Geirhos et al. use a form of statistical analysis called 'error consistency' to compare strategies used by different systems by looking at the kinds of errors they make. Their results show that different DNNs are very consistent in the types of errors they make, providing evidence that they rely on similar strategies. In contrast, when humans and DNNs are compared, their error consistency is only slightly better than what would be predicted by chance alone, providing evidence that they rely on very different strategies. An additional interesting outcome of their analysis is that they also compare CORnet, a recurrent neural network that is considered "most brain-like" according to Brain-Score,[38] on error consistency. They find that CORnet does not do any better than the non-recurrent feedforward convolutional neural networks on error consistency and also that CORnet is very similar to the feedforward convolutional neural networks in terms of error consistency. This implies that even "brain-like" DNNs are still considered to use strategies more similar to other DNNs than to humans. Therefore, if we are interested in the behaviour of DNNs, it is clear there is multiple realization to the extent that they use different strategies from humans.

[38] Brain-Score is a set of neural and behavioural benchmarks used to assess DNNs on their similarity to the brain for object recognition tasks (Schrimpf et al., 2018).

Second, I consider representations – do DNNs and brains use similar representations? Here at first, it appears there is good news for Cao & Yamins' view. Many of the successes attributed to DNN models rely on methods such as representational similarity analysis, which allow comparisons between humans and DNNs to be based on how "similar" the representation matrices are. These results have been used to bolster claims of the similarity between humans and DNNs. However, Mehrer et al. (2020) demonstrate that we cannot guarantee that DNN models can stand in for human internal neural representations. In this paper, the authors show that DNNs of the same network architecture can result in significant individual differences in representations that emerge from the initial random weights of a network. While they recommend some ways to reduce these individual differences, they conclude that there is a strong negative relationship between task performance and representational consistency. Thus, computational neuroscientists should not rely on single networks as models of the brain because the good or bad fit they observe could actually occur due to random weight initialisation instead of an alignment between the network and the brain. Interestingly, Mehrer et al (2020) conclude that scientists should treat DNNs like experimental participants and analyse them in groups instead of focusing on individual models.

By challenging the representational similarity between brains and DNNs, this result poses a problem for Cao & Yamins and indeed any account that seeks to interpret DNNs as providing a monistic mechanistic explanation. This scientific result lines up with philosophical arguments against a monistic view of mechanistic explanation. Hochstein (2016) provides a compelling argument for why individual models rarely provide mechanistic explanations and that instead mechanistic explanation is distributed over sets of scientific models. On his view, mechanistic explanations are often fragmented. They will draw on various features of different models in a set to support inferences to a mechanistic explanation rather than integrating all the models into one

unified model providing a mechanistic explanation. Others have also emphasised the potential for perspectival mechanistic approaches (Dewhurst & Lee, 2021; Kaestner, 2018). I am sympathetic to a more pluralist account of mechanistic explanation.[39] Indeed, I agree with Hochstein that this pluralist account fits better with scientific practice, where scientists view DNN models as experimental participants and organisms (Firestone, 2020; Mehrer et al., 2020).[40] However, I question whether this is sufficient to fulfil the need for wide-scope explanatory features that are present in computational modelling cases. Hochstein's plurality of models seems more in line with the narrow-scope mechanisms that Kaplan describes rather than showing how computational models may inform the wide-scope explanations that are being made in scientific practice. For a wide-scope mechanistic interpretation such as Cao and Yamins' proposal, the evidence that there may be MR at the representational level makes it difficult to defend a model-to-mechanism mapping.

To conclude, I've given two reasons that Cao & Yamins could struggle to support P2 by denying the multiple realization thesis. First, I suggested that there are differences in performance strategies between humans and DNNs, which means we can't deny MRT. Second, I suggested that the lack of representational similarity supports MRT.

[39] As well as pluralist accounts of integration in neuroscience, see Sullivan (2017).

[40] As an aside, I think pluralist accounts could also be compatible with my account of interpreting DNNs as Krogh organisms if these sometimes contribute information and relevant features for a mechanistic explanation.

5.4 Underdetermination argument

So far I have presented evidence that contradicts Cao & Yamins' claims that there is no form of MR in the difficult prediction tasks DNNs perform. In this section, I argue this stems from a more fundamental issue plaguing current mechanistic interpretations of DNNs. Namely, that the success of DNNs does not determine what feature or features are most important for their success. This is an underdetermination argument, whereby the type of biological similarity we can infer between brains and DNNs is underdetermined by the predictive success of DNNs.

While the no-miracles argument Cao and Yamins present is meant to be evidence of the successful identification of the mechanism of the brain, their argument rests on the assumption that this mechanism or the components of the architecture of DNNs is the sole or most important contributor to their success. This is in line with Buckner's (2018) argument that architectural features can explain why deep convolutional neural networks perform so well, which he believes results in them being better suited to depict the generic mechanism responsible for these functions in the brain. Cao & Yamins rely on constraints as a way to support this assumption. For a mechanistic interpretation to hold, you would need to be able to conclude that the performance of DNNs determines facts about the underlying mechanism. However, Weiskopf (2011) argues this reasoning is misguided since all this suggests is that evolution can lead to common functional characteristics and "the presence of a particular functional characteristic does not necessarily entail the presence of any particular physical mechanism" (p.240). One could argue that the situation for DNNs is not as dire as Weiskopf suggests. For example, in chapter 3, I showed that DNN researchers make use of an iterative strategy that attempts to make models and targets more biologically similar. The question still remains though: does the use of an iterative strategy for DNNs guarantee mechanistic similarity? I argue here that, given the current state of DNN research,

it is hard to isolate which features actually account for their predictive success. This is not to categorically rule out that their success is due to mechanistic similarity but here I will provide some examples of alternative 'auxiliary' features that could be responsible instead.

To demonstrate this, let's consider Khaligh-Razavi and Kriegeskorte (2014), a classic example of the predictive success of DNNs that is frequently cited in the "manifestos" of DNN use in neuroscience. In their paper, they show that DNNs can successfully predict and explain neural activity in IT. This seems like unequivocally good news for advocates of DNNs supporting their claim that DNNs are more similar to human and primate representational spaces than conventional vision models. However, the analysis of Khaligh-Razavi and Kriegeskorte (2014) is much more nuanced. The key difference they emphasise between the traditional computer vision models they test and the DNNs is that the latter benefit from intensive supervised training on large, labelled image datasets. This suggests that, in addition to the success of DNNs being due to the potential biological similarity in terms of the mechanism, task training is a crucial factor for the results they acquire. This is further reinforced in Kriegeskorte (2015): "task training of neural networks with millions of labelled images currently provides much stronger constraints than neurophysiological data do on the space of candidate models. Indeed, the recent successes at predicting brain representations of novel natural images are largely driven by task training" (p.17). Khaligh-Razavi and Kriegeskorte (2014) provide several cautions about overinterpreting their results, emphasising the dissimilarities between their DNN, which is a feedforward model, and the visual system, which is recurrent. We should have similar caution in concluding that the success of DNNs is evidence of profound biological similarities with the brain. Since Khaligh-Razavi and Kriegeskorte (2014) are using explicit training and fitting, this study can't tell us why brain regions, such as IT, has the particular representational categories it does (between animate and

inanimate objects) over other options (between human and animal faces). What I take from this is not that we should view the predictive success of DNNs as exclusively telling us about a potential mechanistic mapping but that it tells us something about the significance of potential constraints on realizers, such as tasks and training.

The takeaway is that architectural features alone cannot explain the performance of DNNs. Ultimately, we should also carefully consider the role of the dataset in shaping performance. Other work by the Geirhos lab reinforces this by showing the role that datasets play in how DNNs perform. When Geirhos et al. (2019) found that DNNs relied more on texture rather than shape to categorise objects, they adjusted the dataset to prompt the DNN to rely more on texture. In this case, it would be wrong to use a no-miracles argument to suggest that the success of the DNN is due to architectural similarities without including the way in which the dataset contributes. This is particularly clear when we consider failures of DNNs used in everyday circumstances too. For example, in Buolamwini & Gebru's (2017) survey of commercial image datasets and gender classification systems, they show how the datasets are largely composed of images of lighter-skinned subjects resulting in darker-skinned women being frequently misclassified. Clearly, the dataset needs to be given significant weight in considering DNN performance, a sentiment shared by Khaligh-Razavi and Kriegeskorte (2014). This is also reflected by an increasing interest amongst DNN researchers to develop more ecologically or ethologically valid datasets as a way to improve performance (Cadena et al., 2019; Mehrer et al., 2021).

Another example which serves to illustrate my claim that performance cannot guarantee a particular form of biological similarity is research done using DNNs as models for the human auditory pathway. In Thompson et al. (2021), researchers train convolutional neural networks (CNNs) on auditory recognition tasks and then investigate how they compare to the human

auditory pathway. While the performance of the CNNs correlates with increased representational similarity to fMRI activity in the brain, they do not find evidence of a similar representational hierarchy, as has been shown for DNNs used as models for the visual system. For the auditory case, rather than shallower (or deeper) layers corresponding to earlier (or later) regions, all regions are most similar to one layer of the CNN, the first fully connected layer. The authors suggest their results add nuance to earlier research, such as Kell et al. (2018), that suggests DNNs trained on auditory tasks do replicate a representational hierarchy. In particular, Thompson et al. trained networks for triphone recognition tasks while Kell et al. trained on words. For the purposes of my discussion, it is important that in this case, even though the performance of DNNs is still at human-level, the researchers are aware that this does not guarantee biological similarity in terms of representations. Furthermore, perhaps unsurprisingly, the ability to predict neural activity also does not imply that the models are using the same representations or the same strategies as humans. This reflects a broader issue with the use of benchmarks, which cannot guarantee that performance is achieved in the same way (Evans, Malhotra, & Bowers, 2021). What is apparent from these examples is that decisions about task, training, and datasets will also contribute to important differences between brains and DNNs.

This scientific research highlights the underdetermination of mechanistic similarity by evidence based on performance, which makes it difficult to categorically deny the empirical possibility of multiple realization in these cases. Because of this underdetermination, it is clear we should be exceedingly wary about drawing conclusions from performance to any particular sort of biological similarity, for example about the underlying mechanism, without considering how other factors can be controlled for. This suggests that contra Cao and Yamins' no-miracles argument, a

DNN's success in "difficult" prediction tasks is not enough to secure a model-to-mechanism mapping.

5.5 Responses to the underdetermination argument

In earlier chapters, I have argued that we should view DNNs as providing constraint-based explanations. One might wonder whether constraint-based explanations could still be subsumed under the mechanistic framework. Indeed Cao and Yamins (ms) acknowledge that part of what makes DNNs intelligible is constraint-based reasoning, where constraints "place strong restrictions on the types of possible solutions that will produce it, and what the common characteristics of all such solutions are" (p.17). As I alluded to earlier, they suggest DNNs adhere to two biological constraints – a behavioural constraint based on object recognition performance and an architectural constraint based on the HCNN class of models. Based on the underdetermination argument I have presented here; I don't think the mechanist can go from a no-miracles argument about the need for certain constraints to a claim about mechanistic similarity. It is possible to view constraint-based explanations as providing information about constraints that could then be used in a mechanistic explanation, however, based on the evidence discussed in this paper, that is not currently the case for DNN research. Instead, I think that DNNs are providing constraint-based explanations without the need to commit to mechanistic models.

Consider, for instance, how DNNs can inform us about law-like constraints. Spelling out how DNNs provide a window into law-based constraints is part of making clear how they provide wide-scope explanations. Mechanists already seem to have acknowledged the role of law-like constraints. For example, Stinson (2018) provides several examples of connectionist modellers

discussing constraints in terms of design or general principles with Buckner (2018) agreeing with this analysis. A recent scientific example is Cornford et al. (forthcoming). In this work, they develop Dale's ANNs (DANNs) which are artificial neural networks that adhere to Dale's principle. Dale's principle describes how most neurons have the same functional effect on their post-synaptic neurons. In other words, a neuron can have an either exclusively excitatory or inhibitory effect on post-synaptic cells. The key idea underlying this is that there will be separation of excitation and inhibition in units. Typically, DNNs place no such constraint on their "artificial neurons" which can change weight from positive to negative. One reason for this is because it hinders learning in these networks. Cornford et al. propose adjustments to the learning rule and develop a technique to produce ANNs that adhere to Dale's principle that perform just as well as those that don't adhere to Dale's principle. This use of DNNs can provide scientists with information about the role of law-like constraints even if DNNs fail to be mechanistically similar to the brain.

Finally, one might be concerned whether my underdetermination argument also jeopardises the potential for DNNs to provide constraint-based explanations. If we can't use DNNs' predictive successes to confirm mechanistic similarity, would this cause similar issues when using a DNN's predictive success as evidence for reasoning about law-like constraints on systems? This depends on how strong you take my underdetermination argument to be. If one is pessimistic about the possibility of ever being able to isolate which features are those contributing to the predictive success of DNNs, then this leaves us with a confirmational holism such that we can only test claims about DNNs' performance with respect to all the features and background assumptions that are involved in developing the DNN (Stanford, 2017). Alternatively, you might view this underdetermination as a temporary issue that will be resolved in favour of particular

features as we gain more information about DNNs and the brain. While I won't defend a particular response here, I trust that will satisfy the reader that underdetermination does not need to undermine constraint-based explanations with DNNs.

5.6 Conclusion

This paper sought to challenge certain mechanistic interpretations of DNNs by investigating the role multiple realization plays in assessing mechanistic explanations in computational neuroscience. Despite some philosophers seeing multiple realization as irrelevant to debates on mechanistic explanation, I have argued that recent arguments in favour of DNNs as mechanistic models rely on the denial of MR. By arguing that this assumption does not work for DNNs, I claim that DNNs are subject to an underdetermination argument that makes it difficult to contend that there is no MR involved. Instead, I believe that constraint-based explanation is a better way to assess DNN models in computational neuroscience.

6.0 Concluding remarks

What I have presented in this dissertation scratches the surface of several issues that could be discussed in the philosophy of computational neuroscience and multiple realization. Thus far, I have outlined my account of EMR and some examples of what it might do. For now, let me start by emphasizing two sets of lessons that might be drawn from each chapter and the overarching picture that emerges. Then, I will consider future directions on what EMR could do.

The first set of lessons is more directed at philosophy of computational neuroscience with respect to AI. What I have discussed here has been narrowly focused on DNNs and the role they play in neuroscience practice. However, I do not mean to limit my account to DNNs. If anything, I argue that a perspectival account of multiple realization that admits different roles of biological similarity can help us make sense of the use of AI in cognitive science *tout court*. Boden (2008) suggests that "when connectionists boasted that their models were more biologically plausible than GOFAI, they should rather have said they were *less implausible*" (p.1117). On my account, each approach (and furthermore, each different subfield, lab, or even individual scientist) prioritises its own idea of what counts as relevant similarities between biological systems and machines for the purposes of its research. The utility of a perspectival account of MR is to recognize this and outline a framework going forward.

With this as a starting point, I introduced two strategies a computational neuroscientist could adopt towards using AI for understanding human cognition. The first, an iterative strategy, ultimately aims to reduce multiple realization by working to make the computational model closer to the biological system. In this way, the model will serve as a proxy or surrogate model for the biological system. The second, a comparative strategy, utilizes computational modelling as a way

to maximise multiple realization by working to generate several different realizations of a function. These models can then be used to reason about what constraints are shared amongst realizers and how this affects the realized function.

Both of these strategies come with limitations since the aims of eliminating or maximizing MR are regulative ideals that cannot be fully achieved in practice. That means, in cases of the iterative strategy, we should be cautious to accept claims that ML models will be able to act as complete proxy models, replacing the need for experiments on biological systems, especially just on the basis of selected benchmarks. The ability for AI models to meet particular benchmarks, such as in Schrimpf et al.'s (2018) Brain-Score, does not guarantee that there is deep similarity between how DNNs achieve a function and how the brain does. Similarly, in cases of the comparative strategy, we must be skeptical that ML models will be able to capture all the different ways we would expect variation based on a set of constraints. The methodology we select will itself limit the kind of variation captured in the realizers. If anything, computational neuroscience would do well to consider the comparative strategy more and explore the diversity of realizations of a function. By paying closer attention to the differences between brains and DNNs, this can help avoid overgeneralizing the similarities between artificial and biological systems.

The second set of lessons broadly concerns philosophy of science. In the course of this dissertation, I have established a role that multiple realization can play in philosophy of science. In contrast to some recent concerns that multiple realization can’t really do much for settling philosophy of science debates, I argued that multiple realization provides a framework for constraint-based reasoning. Based on the different conceptions of multiple realization at work in scientific practice, one can establish inferences about constraints on functions or between realizers.

First of all, I view this dissertation as a call for further work on repurposing multiple realization for philosophy of science. While I have provided arguments for how EMR applies to the use of AI in computational neuroscience, my account should not be interpreted as only applying to cases of AI. As I suggest at the end of chapter 3, one can also construe animal models as engineered and apply my account of EMR. Furthermore, I think that other engineering-heavy disciplines such as synthetic biology, applied physics, and computation could be understood through the conceptual framework of EMR.

My account of EMR integrates various threads that are present in the contemporary philosophy of science literature: the relationship between science and engineering, perspectival approaches to science and scientific modelling, and constraint-based reasoning in scientific practice. All of these threads could be expanded upon further.

6.1 Future directions

So far, I have sketched the general lessons and limits to be drawn from this dissertation. However, these ideas can ignite discussion on further issues as well. In closing, I'll consider a few avenues that merit more philosophical meandering.

First, this dissertation has considered the implications of engineered multiple realization for philosophy of science. However, one might ask whether my account of EMR could also have implications for our conceptions of MR in philosophy of mind. One place where I think EMR could lead to the re-appraisal of an existing debate is the question of substrate-independence.

Substrate-independence—the thesis that mental states do not depend on the substrate or matter which realizes them—is a key part of original concept of multiple realizability. As Putnam

remarked, “[w]e could be made of Swiss cheese and it wouldn’t matter” (1975, p.291). Thought experiments rested on similar assumptions, for example, when Pylyshyn (1980) discussed neurons being progressively replaced with chips that are identical in terms of input-output function, the presumption was that you would still go on functioning as before since the material was not a relevant factor in affecting function. This assumption has carried over to the later defenses of computational theories of mind such as Chalmers (1996) as well as being a lynchpin of Bostrom's (2003) simulation argument.

Since EMR is a perspectival account, the assumption of substrate-independence becomes dependent on whether substrate counts as a relevant difference for multiple realization. However, we ought not expect scientific homogeneity on this factor, so not all versions of multiple realization will entail substrate-independence. This contradicts recent accounts of the connection between multiple realization and medium-independence (Ritchie & Piccinini, 2019). On my account, this would require connecting EMR to discussions of implementation, such as Sprevak's (2019). Formulating a response to this would involve considering certain kinds of biological similarity in more detail, namely cellular similarity and metabolic constraints on cognitive processes.

Second, much of contemporary DNN research rests on the assumption that DNNs are involved in “ecologically valid” or “ethologically realistic” tasks (Cao & Yamins, ms.; Yamins & Dicarlo, 2016b). One lesson that can be drawn from my arguments in chapter 3 is the need to re-orient discussions of biological plausibility away from neuro-centric conceptions of what counts as biologically relevant similarities and consider ecological and ethological validity more seriously. My account of DNNs as artificial Krogh organisms implies the importance of utilizing tools for understanding animal cognition in the project of assessing candidate “AI cognitive systems”. There is already a flourishing body of work that looks to draw lessons for the

development of AI from animal cognition research (Buckner, 2019; Crosby, 2020; Firestone, 2020). Building on this, I suggest a fruitful way to further develop this would be to look at discussions of validity of animal models in the philosophical literature, such as Atanasova (2015) and Sullivan (2017).

Third, in this dissertation, I have considered the epistemic factors that contribute to the perspectivism of multiple realization. However, there are also non-epistemic factors involved in considering biologically relevant differences for MR that carry ethical consequences. For example, deciding whether behavioural or neural similarity is important for realizing a cognitive process has implications for assessing animal pain, which could be a significant factor in animal ethics debates. For cases such as AI, I have argued that it is useful for scientists to generate AI systems interpreted as performing cognitive functions, such as object recognition, to better understand biological systems. However, uncritically exporting a single scientific perspective on the biological plausibility of AI systems into social contexts can be harmful. For example, AI systems may be biologically similar enough to inform particular scientific goals such as understanding the relation between tasks and brain function. But this does not necessarily translate to these systems being similar enough to human cognition to replace human counterparts and automate decision-making. To assess the use of AI systems in socio-political contexts would require re-evaluating MR from a different perspective and considering the values that inform and dictate our non-scientific goals, as well as asking what parts of human thinking or decision-making are important for our purposes.

Bibliography

Abraham, T. H. (2019). Cybernetics. In M. Sprevak & M. Colombo (Eds.), *The Routledge handbook of the computational mind* (pp. 52–64). Oxon: Routledge.

Adam, A. (1998). *Artificial knowing: Gender and the thinking machine*. London: Routledge.

Aizawa, K. (2013). Multiple realization by compensatory differences. *European Journal for Philosophy of Science, 3*(1), 69–86. https://doi.org/10.1007/s13194-012-0058-6

Aizawa, K. (2018a). Multiple Realization, Autonomy, and Integration. In D. M. Kaplan (Ed.), *Explanation and Integration in Mind and Brain Science* (pp. 215–235). Oxford: Oxford University Press.

Aizawa, K. (2018b). Multiple realization and multiple "ways" of realization: A progress report. *Studies in History and Philosophy of Science Part A, 68*, 3–9. https://doi.org/10.1016/j.shpsa.2017.11.005

Aizawa, K. (2020). The many problems of multiple realization. *American Philosophical Quarterly, 57*(1), 3–16.

Aizawa, K., & Gillett, C. (2009). The (multiple) realization of psychological and other properties in the sciences. *Mind and Language*, *24*(2), 181–208. https://doi.org/10.1111/j.1468-0017.2008.01359.x

Amundson, R. (2000). Against normal function. *Studies in History and Philosophy of Science Part C :Studies in History and Philosophy of Biological and Biomedical Sciences*, *31*(1), 33–53. https://doi.org/10.1016/S1369-8486(99)00033-3

Anderson, M. L. (2015). Beyond Componential Constitution in the Brain: Starburst Amacrine Cells and Enabling Constraints. *Open MIND*, *1*, 1–13. https://doi.org/10.15502/9783958570429

Ankeny, R. A., & Leonelli, S. (2011). What is So Special About Model Organisms? *Studies in History and Philosophy of Science Part A, 42*(2), 313–323.

Ankeny, R. A., & Leonelli, S. (2020). *Model Organisms*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108593014

Atanasova, N. (2015). Validating Animal Models. *Theoria*, *30*(2), 163–181. https://doi.org/10.1387/theoria.12761

Balari, S., & Lorenzo, G. (2019). Realization in biology? *History and Philosophy of the Life Sciences*, *41*(1), 1–27. https://doi.org/10.1007/s40656-019-0243-4

Bantegnie, B. (2017). Commentary: Integrative modeling and the role of neural constraints. *Frontiers in Psychology*, *8*. https://doi.org/10.1086/687854

Batterman, R. W. (2000). Multiple Realizability and Universality. *British Journal for the Philosophy of Science*, *51*(1), 115–145. https://doi.org/10.1093/bjps/51.1.115

Batterman, R. W., & Rice, C. C. (2014). Minimal model explanations. *Philosophy of Science*, *81*(3), 349–376. https://doi.org/10.1086/676677

Bechtel, W., & Abrahamsen, A. (2002). *Connectionism and the Mind: Parallel Processing, Dynamics, and Evolution in Networks* (2nd ed.). Oxford: Blackwell Publishers Ltd.

Bechtel, W., & Mundale, J. (1999). Multiple Realizability Revisited: Linking Cognitive and Neural States. *Philosophy of Science*, *66*(2), 175–207.

Bengio, Y., & LeCun, Y. (2007). Scaling Learning Algorithms toward AI. *Large-Scale Kernel Machines*, *34*(5), 1–41. https://doi.org/10.7551/mitpress/7496.003.0016

Bengio, Y., Lee, D., Bornschein, J., Mesnard, T., & Lin, Z. (2015). Towards Biologically Plausible Deep Learning. *ArXiv Preprint ArXiv:1502.04156*.

Beyret, B., Hernández-Orallo, J., Cheke, L., Halina, M., Shanahan, M., & Crosby, M. (2019). The Animal-AI Environment: Training and Testing Animal-Like Artificial Cognition. *ArXiv Preprint ArXiv:1909.07483*.

Boden, M. A. (2008). *Mind as machine: A history of cognitive science*. Oxford: Oxford University Press.

Bolker, J. A. (2009). Exemplary and surrogate models: Two modes of representation in biology. *Perspectives in Biology and Medicine*, *52*(4), 485–499.

Bostrom, N. (2003). Are we living in a computer simulation? *The Philosophical Quarterly*, *53*(211), 243–255.

Buckner, C. (2018). Empiricism without Magic: Transformational Abstraction in Deep Convolutional Neural Networks. *Synthese*, *195*(12), 5339–5372.

Buckner, C. (2019). *The Comparative Psychology of Artificial Intelligence*. *Preprint*. Retrieved from http://philsci-archive.pitt.edu/id/eprint/16128 (accessed 2021-03-16)

Buolamwini, J., & Gebru, T. (2017). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *OncoTargets and Therapy*, *10*, 1889–1896. https://doi.org/10.2147/OTT.S126905

Cadena, S. A., Sinz, F. H., Muhammad, T., Froudarakis, E., Cobos, E., Walker, E. Y., … Ecker, A. S. (2019). How well do deep neural networks trained on object recognition characterize the mouse visual system? *NeurIPS Neuro AI Workshop 2019*, 1–5.

Calisi, R. M., & Bentley, G. E. (2009). Lab and field experiments: Are they the same animal? *Hormones and Behavior*, *56*(1), 1–10. https://doi.org/10.1016/j.yhbeh.2009.02.010

Canguilhem, G. (1963). The role of analogies and models in biological discovery. In A. C. Crombie (Ed.), *Scientific Change: Historical studies in the intellectual, social and technical conditions for scientific discovery and technical invention, from antiquity to the present* (pp. 507–520). London: Heinemann.

Canguilhem, G. (1978). *The Normal and the Pathological* (1st ed.). Dordrecht, Holland: D. Reidel Publishing Company.

Cao, R., & Yamins, D. (n.d.). Making sense of mechanism : How neural network models can explain brain function, 1–33.

Carandini, M. (2012). From circuits to behavior: A bridge too far? *Nature Neuroscience*, *15*(4), 507–509. https://doi.org/10.1038/nn.3043

Carandini, M., & Heeger, D. J. (2012). Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, *13*(1), 51–62. https://doi.org/10.1038/nrn3136

Caron, S., & Abbott, L. F. (2017). Neuroscience: Intelligence in the Honeybee Mushroom Body. *Current Biology*, *27*(6), R220–R223. https://doi.org/10.1016/j.cub.2017.02.011

Cave, S., & Dihal, K. (2018). Ancient dreams of intelligent machines: 3,000 years of robots. *Nature*, *559*(7715), 473–475.

Chalmers, D. (1996). Absent Qualia, Fading Qualia, Dancing Qualia. In *The conscious mind: In search of a fundamental theory*. New York: Oxford University Press.

Chin-Yee, B., & Upshur, R. (2019). Three problems with big data and artificial intelligence in medicine. *Perspectives in Biology and Medicine*, *62*(2), 237–256. https://doi.org/10.1353/pbm.2019.0012

Chirimuuta, M. (2014). Minimal models and canonical neural computations: the distinctness of computational explanation in neuroscience. *Synthese*, *191*(2), 127–153.

Chirimuuta, M. (2018a). Explanation in computational neuroscience: Causal and non-causal. *British Journal for the Philosophy of Science*, *69*(3), 849–880. https://doi.org/10.1093/bjps/axw034

Chirimuuta, M. (2018b). Marr, Mayr, and MR: What functionalism should now be about. *Philosophical Psychology*, *31*(3), 403–418.

Chirimuuta, M. (2020a). Charting the Heraclitean Brain: Perspectivism and Simplification in Models of the Motor Cortex. In M. Massimi & C. D. McCoy (Eds.), *Understanding Perspectivism: Scientific Challenges and Methodological Prospects* (pp. 141–159). New York: Routledge.

Chirimuuta, M. (2020b). Your Brain is Like a Computer: Function, Analogy, Simplification. In F. Calzavarini & M. Viola (Eds.), *Neural mechanisms: New challenges in the philosophy of neuroscience*. Berlin: Springer.

Chittka, L., & Niven, J. (2009). Are Bigger Brains Better? *Current Biology*, *19*(21), R995–R1008. https://doi.org/10.1016/j.cub.2009.08.023

Cichy, R. M., & Kaiser, D. (2019). Deep Neural Networks as Scientific Models. *Trends in Cognitive Sciences*, *23*(4), 305–317.

Cornford, J., Kalajdzievski, D., Leite, M., Lamarquette, A., Kullmann, D. M., & Richards, B. A. (n.d.). Learning to live with Dale's principle: ANNs with separate excitatory and inhibitory units. In *ICLR 2021* (pp. 1–12).

Craver, C. F. (2007). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. Oxford: Oxford University Press.

Crick, F. (1989). The recent excitement about neural networks. *Nature*, *337*(6203), 129–132. https://doi.org/10.1038/337129a0

Crosby, M. (2020). Building Thinking Machines by Solving Animal Cognition Tasks. *Minds and Machines*, *30*(4), 589–615. https://doi.org/10.1007/s11023-020-09535-6

Crosby, M., Beyret, B., & Halina, M. (2019). The Animal-AI Olympics. *Nature Machine Intelligence*, *1*(5), 257–257.

Danks, D. (2020). Safe-and-Substantive Perspectivism. In M. Massimi & C. D. McCoy (Eds.), *Understanding Perspectivism: Scientific Challenges and Methodological Prospects* (pp. 127–140). New York: Routledge.

Delahunt, C. B., Riffell, J. A., & Kutz, J. N. (2018). Biological Mechanisms for Learning: A Computational Model of Olfactory Learning in the Manduca sexta Moth, with Applications to Neural Nets, *12*(December), 1–20. https://doi.org/10.3389/fncom.2018.00102

Dennett, D. C. (1978). Why You Can't Make a Computer That Feels Pain. *Synthese*, *38*(3), 415–456.

Dewhurst, J., & Lee, J. (2021). The mechanistic stance. *European Journal for Philosophy of Science*, *11*(1), 1–21.

Dick, S. (2015). Of models and machines: Implementing bounded rationality. *Isis*, *106*(3), 623–634. https://doi.org/10.1086/683527

Dietrich, M. R., Ankeny, R. A., Crowe, N., Green, S., & Leonelli, S. (2019). How to choose your research organism. *Studies in History and Philosophy of Biological and Biomedical Sciences*. https://doi.org/10.1016/j.shpsc.2019.101227

Ensmenger, N. (2012). Is chess the drosophila of artificial intelligence? A social history of an

algorithm. *Social Studies of Science*, *42*(1), 5–30.

Evans, B. D., Malhotra, G., & Bowers, J. S. (2021). Biological convolutions improve DNN robustness to noise and generalisation. *BioRxiv*, 2021.02.18.431827. https://doi.org/10.1101/2021.02.18.431827

Figdor, C. (2010). Neuroscience and the multiple realization of cognitive functions. *Philosophy of Science*, *77*(3), 419–456.

Firestone, C. (2020). Performance vs. competence in human–machine comparisons. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(43), 26562–26571. https://doi.org/10.1073/pnas.1905334117

Fodor, J. A., & Block, N. (1972). What Psychological States are Not. *The Philosophical Review*, *81*(2), 159–181. https://doi.org/10.2307/2176743

Franchi, S., & Güzeldere, G. (2005). Machinations of the Mind: Cybernetics and Artificial Intelligence from Automata to Cyborgs. In S. Franchi & G. Güzeldere (Eds.), *Mechanical bodies, computational minds: artificial intelligence from automata to cyborgs* (pp. 15–152). Cambridge, MA: MIT Press.

Franklin-Hall, L. R. (2016). New Mechanistic Explanation and the Need for Explanatory Constraints. In *Scientific Composition and Metaphysical Ground* (pp. 41–74). London: Palgrave Macmillan.

Fukushima, K. (1980). Neocognitron: a self-organising Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position. *Biol. Cybernetics*, *36*, 193–202. https://doi.org/10.1007/BF00344251

Geirhos, R., Meding, K., & Wichmann, F. A. (2020). Beyond accuracy: quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency. *ArXiv*, 1–21. Retrieved from http://arxiv.org/abs/2006.16736

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2019). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *ICLR 2019* (pp. 1–22).

Geisler, W. S. (2001). Ideal Observer Analysis. In *The Visual Neurosciences* (Vol. 10, pp. 244–246). Cambridge, MA: MIT Press.

Giere, R. N. (2010). *Scientific perspectivism*. Chicago: University of Chicago Press.

Glaser, J. I., Benjamin, A. S., Farhoodi, R., & Kording, K. P. (2019). The roles of supervised machine learning in systems neuroscience. *Progress in Neurobiology*, *175*, 126–137. https://doi.org/10.1016/j.pneurobio.2019.01.008

Godfrey-smith, P. (2008). Reduction in Real Life. In J. Hohwy & J. Kallestrup (Eds.), *Being Reduced: New Essays on Causation and Explanation in the Special Sciences* (pp. 52–74).

Oxford: Oxford University Press.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and Harnessing Adversarial Examples. *ArXiv*, 1–11. Retrieved from http://arxiv.org/abs/1412.6572

Green, S., Dietrich, M. R., Leonelli, S., & Ankeny, R. A. (2018). 'Extreme' organisms and the problem of generalization: interpreting the Krogh principle. *History and Philosophy of the Life Sciences*, *40*(65), 1–22.

Green, S., & Jones, N. (2016). Constraint-Based Reasoning for Search and Explanation: Strategies for Understanding Variation and Patterns in Biology. *Dialectica*, *70*(3), 343–374.

Gregory, R. L. (1961). The brain as an engineering problem. In *Current problems in animal behaviour* (pp. 307–330).

Guest, O., & Love, B. C. (2019). Levels of Representation in a Deep Learning Model of Categorization. *BioRxiv*, 626374. https://doi.org/10.1101/626374

Haesemeyer, M., Schier, A. F., & Engert, F. (2019). Convergent Temperature Representations in Artificial and Biological Neural Networks. *Neuron*, *103*(6), 1123-1134.e6. https://doi.org/10.1016/j.neuron.2019.07.003

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-Inspired Artificial Intelligence. *Neuron*, *95*(2), 245–258. https://doi.org/10.1016/j.neuron.2017.06.011

Hasson, U., Nastase, S. A., & Goldstein, A. (2020). Direct Fit to Nature: An Evolutionary Perspective on Biological and Artificial Neural Networks. *Neuron*, *105*(3), 416–434. https://doi.org/10.1016/j.neuron.2019.12.002

Haugeland, J. (1989). *Artificial Intelligence: The Very Idea*. Cambridge, MA: MIT Press.

Hochstein, E. (2016). One mechanism, many models: a distributed theory of mechanistic explanation. *Synthese*, *193*(5), 1387–1407. https://doi.org/10.1007/s11229-015-0844-8

Hofstadter, D. R. (1985). Waking up from the Boolean dream, or, subcognition as computation. In *Metamagical themas: Questing for the essence of mind and pattern* (pp. 631–665). New York: Basic Books.

Holland, O. (2003). The first biologically inspired robots. *Robotica*, *21*(4), 351–363. https://doi.org/10.1017/S0263574703004971

Illari, P. (2013). Mechanistic explanation: Integrating the ontic and epistemic. *Erkenntnis*, *78*(2), 237–255.

Isaac, A. M. C. (2019). Computational thought from Descartes to Lovelace. In M. Sprevak & M. Colombo (Eds.), *The Routledge handbook of the computational mind* (pp. 9–22). Oxon: Routledge.

Izquierdo, E. J., & Beer, R. D. (2013). Connecting a Connectome to Behavior: An Ensemble of Neuroanatomical Models of C. elegans Klinotaxis. *PLoS Computational Biology*, *9*(2), 1–20.

Jonas, E., & Kording, K. P. (2017). Could a Neuroscientist Understand a Microprocessor ? *PLoS Computational Biology*, *13*(1), 1–24. https://doi.org/10.1371/journal.pcbi.1005268

Kaestner, L. (2018). Integrating mechanistic explanations through epistemic perspectives. *Studies in History and Philosophy of Science Part A*, *68*, 68–79.

Kaplan, D. M. (2011). Explanation and description in computational neuroscience. *Synthese*, *183*(3), 339–373. https://doi.org/10.1007/sl

Kaplan, D. M. (2017). Neural computation, multiple realizability, and the prospects for mechanistic explanation. In D. M. Kaplan (Ed.), *Explanation and Integration in Mind and Brain Science* (pp. 164–189). Oxford: Oxford University Press.

Kaplan, D. M., & Craver, C. F. (2011). The explanatory force of dynamical and mathematical models in neuroscience: A mechanistic perspective. *Philosophy of Science*, *78*(4), 601–627. https://doi.org/10.1086/661755

Kawamura, A., Kohri, M., Morimoto, G., Nannichi, Y., Taniguchi, T., & Kishikawa, K. (2016). Full-Color Biomimetic Photonic Materials with Iridescent and Non-Iridescent Structural Colors. *Scientific Reports*, *6*, 1–10. https://doi.org/10.1038/srep33984

Kay, K. N. (2018). Principles for models of neural information processing. *NeuroImage*, *180*, 101–109. https://doi.org/10.1016/j.neuroimage.2017.08.016

Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018). A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron*, *98*(3), 630-644.e16. https://doi.org/10.1016/j.neuron.2018.03.044

Kell, A. J., & McDermott, J. H. (2019). Deep neural network models of sensory systems: windows onto the role of task constraints. *Current Opinion in Neurobiology*, *55*, 121–132. https://doi.org/10.1016/j.conb.2019.02.003

Kelty, C. M. (2012). This is not an article: Model organism newsletters and the question of "open science." *BioSocieties*, *7*(2), 140–168. https://doi.org/10.1057/biosoc.2012.8

Key, B. (2015). Fish do not feel pain and its implications for understanding phenomenal consciousness. *Biology and Philosophy*, *30*(2), 149–165. https://doi.org/10.1007/s10539-014-9469-4

Khaligh-Razavi, S. M., & Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLoS Computational Biology*, *10*(11). https://doi.org/10.1371/journal.pcbi.1003915

Kingstone, A., Smilek, D., & Eastwood, J. D. (2008). Cognitive Ethology: A new approach for

studying human cognition. *British Journal of Psychology*, *99*(3), 317–340. https://doi.org/10.1348/000712607X251243

Kirk, R. G. W. (2012). "Standardization through Mechanization": Germ-free Life and the Engineering of the Ideal Laboratory Animal. *Technology and Culture*, *53*(1), 61–93.

Klein, C. (2013). Multiple realizability and the semantic view of theories. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *163*(3), 683–695.

Kohler, R. E. (1994). *Lords of the fly: Drosophila genetics and the experimental life*. Chicago: University of Chicago Press.

Koskinen, R. (2019). Multiple Realizability as a design heuristic in biological engineering. *European Journal for Philosophy of Science*, *9*(15), 1–15.

Kriegeskorte, N. (2015). Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annual Review of Vision Science*, *1*, 417–446.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*(4). https://doi.org/10.3389/neuro.06.004.2008

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 1097–1105.

Krogh, A. (1929). The Progress of Physiology. *The American Journal of Physiology*, *90*(2), 243–251.

Kubilius, J., Bracci, S., & Op de Beeck, H. P. (2016). Deep Neural Networks as a Computational Model for Human Shape Sensitivity. *PLoS Computational Biology*, *12*(4), 1–26.

Kurakin, A., Goodfellow, I. J., & Bengio, S. (2018). Adversarial examples in the physical world. In *Artificial Intelligence Safety and Security* (pp. 99–112). Boca Raton: CRC Press.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building Machines That Learn and Think Like People. *Behavioral and Brain Sciences*, *40*, 1–101. https://doi.org/10.1017/S0140525X16001837

Larkin, A., Marygold, S. J., Antonazzo, G., Attrill, H., dos Santos, G., Garapati, P. V, … Flybase Consortium, . (2021). Flybase: updates to the Drosophila melanogaster knowledge base. *Nucleic Acids Research: Database Issue*, *49*(D1), D899–D907.

Leonelli, S. (2007). Growing weed, producing knowledge: an epistemological history of Arabidopsis thaliana. *History and Philosophy of the Life Sciences*, *29*(2), 193–223.

Leonelli, S., & Ankeny, R. A. (2013). What makes a model organism? *Endeavour*, *37*(4), 209–212.

Leonelli, S., Ankeny, R. A., Nelson, N. C., & Ramsden, E. (2014). Making Organisms Model Human Behavior : Situated Models in North-American Alcohol Research, 1950-onwards. *Sci Context*, *27*(3), 485–509.

Lettvin, J. Y. (2000). Jerome Y. Lettvin. In J. A. Anderson & E. Rosenfeld (Eds.), *Talking Nets: An oral history of neural networks*. Cambridge, MA: MIT Press.

Levy, A. (2014). What was Hodgkin and Huxley's achievement? *British Journal for the Philosophy of Science*, *65*(3), 469–492.

Levy, A., & Currie, A. (2015). Model organisms are not (theoretical) models. *British Journal for the Philosophy of Science*, *66*(2), 327–348.

Lillicrap, T. P., Cownden, D., Tweed, D. B., & Akerman, C. J. (2016). Random synaptic feedback weights support error backpropagation for deep learning. *Nature Communications*, *7*(1), 1–10. https://doi.org/10.1038/ncomms13276

Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J., & Hinton, G. (2020). Backpropagation and the brain. *Nature Reviews Neuroscience*, *21*(6), 335–346. https://doi.org/10.1038/s41583-020-0277-3

Lindsay, G. W. (2020). Convolutional Neural Networks as a Model of the Visual System: Past, Present, and Future. *Journal of Cognitive Neuroscience*, 1–15.

Lindsay, G. W., & Miller, K. D. (2017). Understanding Biological Visual Attention Using Convolutional Neural Networks. *BioRxiv*, 233338. https://doi.org/10.1101/233338

López-Rubio, E. (2018). Computational Functionalism for the Deep Learning Era. *Minds and Machines*, *28*(4), 667–688. https://doi.org/10.1007/s11023-018-9480-7

Luce, R. D. (1995). Four tensions concerning mathematical modeling in psychology. *Annual Review of Psychology*, *46*, 1–26.

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about Mechanisms. *Philosophy of Science*, *67*(1), 1–25. https://doi.org/10.1086/392759

Marblestone, A. H., Wayne, G., Kording, K. P., & Scholte, H. S. (2016). Toward an Integration of Deep Learning and Neuroscience. *Frontiers in Computational Neuroscience*, *10*, 1–41. https://doi.org/10.3389/fncom.2016.00094

Massimi, M. (2017). Perspectivism. In J. Saatsi (Ed.), *The Routledge Handbook of Scientific Realism* (pp. 164–175). Oxon: Routledge.

Mayor, A. (2018). *Gods and Robots: Myths, Machines, and Ancient Dreams of Technology*. Princeton: Princeton University Press.

McCarthy, J. (1990). Chess as the Drosophila of AI. In T. A. Marsland & J. Schaeffer (Eds.), *Computers, Chess, and Cognition* (pp. 227–237). New York: Springer.

https://doi.org/10.1007/978-1-4613-9080-0_14

McClelland, J. L., Rumelhart, D. E., & PDP Research Group. (1987). *Parallel Distributed Processing, Vol.2: Psychological and Biological Models*. MIT Press.

Medler, D. A. (1998). A Brief History of Connectionism. *Neural Computing Surveys*, *1*(2), 18–73.

Mehrer, J., Spoerer, C. J., Jones, E. C., Kriegeskorte, N., & Kietzmann, T. C. (2021). An ecologically motivated image dataset for deep learning yields better models of human vision. *Proceedings of the National Academy of Sciences*, *118*(8), e2011417118. https://doi.org/10.1073/pnas.2011417118

Mehrer, J., Spoerer, C. J., Kriegeskorte, N., & Kietzmann, T. C. (2020). Individual differences among deep neural network models. *Nature Communications*, *11*(1), 1–12. https://doi.org/10.1038/s41467-020-19632-w

Menzel, R. (2012). The honeybee as a model for understanding the basis of cognition. *Nature Reviews. Neuroscience*, *13*(11), 758–768. https://doi.org/10.1038/nrn3357

Meunier, R. (2012). Stages in the development of a model organism as a platform for mechanistic models in developmental biology: Zebrafish, 1970-2000. *Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*(2), 522–531. https://doi.org/10.1016/j.shpsc.2011.11.013

Michel, M. (2019). Fish and Microchips: on fish pain and multiple realization. *Philosophical Studies*, *176*(9), 2411–2428. https://doi.org/10.1007/s11098-018-1133-4

Milani, L., & Ghiselli, F. (2019). Faraway, so close. The comparative method and the potential of non-model animals in mitochondrial research. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1790). https://doi.org/10.1098/rstb.2019.0186

Milkowski, M. (2013). *Explaining the computational mind*. Cambridge, MA: MIT Press.

Miłkowski, M. (2018). Embodied Cognition Meets Multiple Realizability. *Reti, Saperi, Linguaggi*, (2), 349–364. https://doi.org/10.12832/92305

Millikan, R. G. (1999). Historical Kinds and the "Special Sciences." *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *95*(1/2), 45–65.

Mitchell, S. D. (2020). Perspectives, Representation, and Integration. In M. Massimi & C. D. McCoy (Eds.), *Understanding Perspectivism: Scientific Challenges and Methodological Prospects*. New York: Routledge.

Mitchinson, B., Pearson, M. J., Pipe, A. G., & Prescott, T. J. (2011). Biomimetic robots as scientific models: A view from the whisker tip. In J. L. Krichmar & H. Wagatsuma (Eds.), *Neuromorphic and Brain-Based Robots* (pp. 23–57). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511994838.004

Morrison, M., & Morgan, M. S. (1999). Models as Mediating Instruments. In M. S. Morgan & M. Morrison (Eds.), *Models as mediators: Perspectives on natural and social science* (pp. 10–37). Cambridge: Cambridge University Press.

Musall, S., Urai, A. E., Sussillo, D., & Churchland, A. K. (2019). Harnessing behavioral diversity to understand neural computations for cognition. *Current Opinion in Neurobiology*, *58*, 229–238.

Naselaris, T., Bassett, D. S., Fletcher, A. K., Kording, K., Kriegeskorte, N., Nienborg, H., … Kay, K. (2018). Cognitive Computational Neuroscience: A New Conference for an Emerging Discipline. *Trends in Cognitive Sciences*, *22*(5), 365–367.

National Research Council, . (1999). Marine Organisms as Models for Biomedical Research. In *From Monsoons to Microbes: Understanding the Ocean's Role in Human Health* (pp. 83–96). Washington, DC: The National Academies Press. https://doi.org/10.17226/6368

Otis, L. (2001). The Other End of the Wire: Uncertainties of Organic and Telegraphic Communication. *Configurations*, *9*(2), 181–206.

Parker, G. A., & Smith, J. M. (1990). Optimality theory in evolutionary biology. *Nature*, *348*(6296), 27–33.

Parkkinen, V. P. (2017). Are model organisms theoretical models? *Disputatio*, *9*(47), 471–498. https://doi.org/10.1515/disp-2017-0015

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2018). Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive Science*, *42*(8), 2648–2669.

Piccinini, G. (2004). The First Computational Theory of Mind and Brain: A Close Look at McCulloch and Pitts's "Logical Calculus of Ideas Immanent in Nervous Activity." *Synthese*, *141*, 175–215.

Piccinini, G. (2015). *Physical computation: A mechanistic account*. Oxford: Oxford University Press.

Piccinini, G. (2020). *Neurocognitive Mechanisms: Explaining Biological Cognition*. Oxford: Oxford University Press.

Piccinini, G., & Maley, C. J. (2014). The Metaphysics of Mind and the multiple sources of multiple realizability. In M. Sprevak & J. Kallestrup (Eds.), *New Waves in Philosophy of Mind* (pp. 125–152). London: Palgrave Macmillan.

Polger, T. W., & Shapiro, L. A. (2016). *The multiple realization book*. Oxford: Oxford University Press.

Potochnik, A. (2012). Feminist implications of model-based science. *Studies in History and Philosophy of Science Part A*, *43*(2), 383–389. https://doi.org/10.1016/j.shpsa.2011.12.033

Potochnik, A., & McGill, B. (2012). The limitations of hierarchical organization. *Philosophy of Science*, *79*, 120–140. https://doi.org/10.1086/663237

Prabhu, V. U., & Birhane, A. (2021). *Large Image Datasets: A Pyrrhic Win for Computer Vision? Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*.

Prescott, T. J., Pearson, M. J., Mitchinson, B., Sullivan, J. C. W., & Pipe, A. G. (2009). Whisking with robots: From rat vibrissae to biomimetic technology for active touch. *IEEE Robotics and Automation Magazine*, *16*(3), 42–50. https://doi.org/10.1109/MRA.2009.933624

Preston, B. (2009). Philosophical theories of artifact function. In A. Meijers (Ed.), *Philosophy of technology and engineering sciences* (pp. 213–233). Burlington, MA: Elsevier.

Putnam, H. (1964). Robots: Machines or Artificially Created Life? *The Journal of Philosophy*, *61*(21), 668–691.

Putnam, H. (1967). Psychological Predicates. In W. H. Capitan & D. D. Merrill (Eds.), *Art, mind and religion*. Pittsburgh: University of Pittsburgh Press.

Putnam, H. (1975). Philosophy and our mental life. In *Mind, Language and Reality*. Cambridge: Cambridge University Press.

Pylyshyn, Z. W. (1980). The "causal power" of machines. *Behavioral and Brain Sciences*, 442–444.

Rader, K. (2004). *Making mice: Standardizing animals for American biomedical research, 1900-1955*. Princeton: Princeton University Press.

Raja, V., & Anderson, M. L. (2020). Behavior considered as an enabling constraint. In F Calzavarini & M. Viola (Eds.), *Neural mechanisms: New challenges in the philosophy of neuroscience* (pp. 209–232). New York: Springer.

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., … Kording, K. P. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, *22*(11), 1761–1770.

Riskin, J. (2007). Eighteenth-Century Wetware. In B. Bensaude-Vincent & W. R. Newman (Eds.), *The Artificial and the Natural* (pp. 239–274). Cambridge, MA: MIT Press.

Riskin, J. (2016). *The restless clock: A history of the centuries-long argument over what makes living things tick*. Chicago: University of Chicago Press.

Ritchie, J. B., & Piccinini, G. (2019). Computational implementation. In M. Sprevak & M. Colombo (Eds.), *The Routledge handbook of the computational mind* (pp. 192–204). Oxon: Routledge.

Ritter, S., Barrett, D. G. T., Santoro, A., & Botvinick, M. M. (2017). Cognitive Psychology for Deep Neural Networks: A Shape Bias Case Study. In *Proceedings of the 34th International*

Conference on Machine Learning. Sydney, Australia. https://doi.org/10.1037/a0037840

Roper, M., Fernando, C., & Chittka, L. (2017). Insect Bio-inspired Neural Network Provides New Evidence on How Simple Feature Detectors Can Enable Complex Visual Generalization and Stimulus Location Invariance in the Miniature Brain of Honeybees. *PLoS Computational Biology*, *13*(2), 1–23. https://doi.org/10.1371/journal.pcbi.1005333

Rosenbleuth, A., & Wiener, N. (1945). The Role of Models in Science. *Philosophy of Science*, *12*(4), 316–321.

Ross, L. N. (n.d.). The explanatory nature of constraints: Law-based, mathematical, and causal, 1–17.

Russell, J. J., Theriot, J. A., Sood, P., Marshall, W. F., Landweber, L. F., Fritz-Laylin, L., … Brunet, A. (2017). *Non-model model organisms*. *BMC Biology* (Vol. 15). BMC Biology. https://doi.org/10.1186/s12915-017-0391-5

Sabour, S., Frosst, N., & Hinton, G. E. (2017). Dynamic Routing Between Capsules. *ArXiv Preprint*. Retrieved from http://arxiv.org/abs/1710.09829

Saito, A. (2011). Material design and structural color inspired by biomimetic approach. *Science and Technology of Advanced Materials*, *12*(6), 064709. https://doi.org/10.1088/1468-6996/12/6/064709

Scholte, H. S. (2018). Fantastic DNimals and where to find them. *NeuroImage*, *180*, 112–113.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., … DiCarlo, J. J. (2018). Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? *BioRxiv*, *407007*. https://doi.org/10.1101/407007

Sejnowski, T. J., Koch, C., & Churchland, P. S. (1988). Computational Neuroscience. *Science*, *241*(4871), 1299–1306.

Shapiro, L. (2004). *The Mind Incarnate*. Cambridge, MA: MIT Press.

Sherrington, C. (1940). *Man on his Nature*. Cambridge: Cambridge University Press.

Simon, H. A., & Newell, A. (1964). Information Processing in Computer and Man. *American Scientist*, *52*(3), 281–300.

Sneddon, L. U., Elwood, R. W., Adamo, S. A., & Leach, M. C. (2014). Defining and assessing animal pain. *Animal Behaviour*, *97*, 201–212. https://doi.org/10.1016/j.anbehav.2014.09.007

Sober, E. (2010). Learning from functionalism: Prospects for strong artificial life. *The Nature of Life: Classical and Contemporary Perspectives from Philosophy and Science*, (January 1991), 225–235. https://doi.org/10.1017/CBO9780511730191.021

Sprevak, M. (2019). Triviality arguments against implementation. In *The Routledge handbook of*

the computational mind (pp. 175–191). Oxon: Routledge.

Stanford, K. (2017). Underdetermination of Scientific Theory. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy (Winter 2017 Edition)*.

Sterling, P., & Laughlin, S. (2015). *Principles of Neural Design*. Cambridge, MA: MIT Press.

Stinson, C. (2018). Explanation and Connectionist Models. In M. Sprevak & M. Colombo (Eds.), *The Routledge handbook of the computational mind*. New York: Routledge.

Stinson, C. (2020). From Implausible Artificial Neurons to Idealized Cognitive Models: Rebooting Philosophy of Artificial Intelligence. *Philosophy of Science*, 1–38. https://doi.org/https://doi.org/10.1086/709730

Sullivan, J. A. (2017). Coordinated pluralism as a means to facilitate integrative taxonomies of cognition. *Philosophical Explorations*, *20*(2), 129–145.

Tarr, M. J., & Aminoff, E. M. (2016). Can big data help us understand human vision? In M. N. Jones (Ed.), *Big Data in Cognitive Science* (pp. 343–363). New York: Routledge.

Thompson, J. A. F., Bengio, Y., Formisano, E., & Schönwiesner, M. (2021). Training neural networks to recognize speech increased their correspondence to the human auditory pathway but did not yield a shared hierarchy of acoustic features. Retrieved from https://doi.org/10.1101/2021.01.26.428323

Towl, B. (2012). Laws and constrained kinds: a lesson from motor neuroscience. *Synthese*, *189*(3), 433–450.

Van Fraassen, B. C. (2008). *Scientific Representation*. Oxford: Oxford University Press.

Weisberg, M. (2012). *Simulation and similarity: using models to understand the world*. Oxford: Oxford University Press.

Weiskopf, D. A. (2011). The functional unity of special science kinds. *British Journal for the Philosophy of Science*, *62*(2), 233–258. https://doi.org/10.1093/bjps/axq026

Weiskopf, D. A. (2016). Integrative modeling and the role of neural constraints. *Philosophy of Science*, *83*(5), 674–685. https://doi.org/10.1086/687854

Wimsatt, W. C. (1994). The Ontology of Complex Systems: Levels of Organization, Perspectives, and Causal Thickets. *Canadian Journal of Philosophy*, *20*, 207–274.

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, *1*(1), 67–82. https://doi.org/10.1109/4235.585893

Yamins, D. L. K., & Dicarlo, J. J. (2016). Eight open questions in the computational modeling of higher sensory cortex. *Current Opinion in Neurobiology*, *37*, 114–120.

https://doi.org/10.1016/j.conb.2016.02.001

Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356–365. https://doi.org/10.1038/nn.4244

Yang, Z., Zhou, J., Wei, B., Cheng, Y., Zhang, L., & Zhen, X. (2019). Comparative transcriptome analysis reveals osmotic-regulated genes in the gill of Chinese mitten crab (Eriocheir sinensis). *PLoS ONE*, *14*(1), 1–16. https://doi.org/10.1371/journal.pone.0210469

Yartsev, M. M. (2017). The emperor's new wardrobe: Rebalancing diversity of animal models in neuroscience research. *Science*, *358*(6362), 466–469. https://doi.org/10.1126/science.aan8865